5-2011

A Comparison of Spatial Prediction Techniques Using Both Hard and Soft Data

Megan L. Liedtke Tesar
University of Nebraska-Lincoln, megan.liedtke@doane.edu

A COMPARISON OF SPATIAL PREDICTION TECHNIQUES USING BOTH HARD AND SOFT DATA

by

Megan Lynne Liedtke Tesar

A DISSERTATION

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Doctor of Philosophy

Major: Statistics

Under the Supervision of Professor David B. Marx

Lincoln, Nebraska

May, 2011

A COMPARISON OF SPATIAL PREDICTION TECHNIQUES USING BOTH HARD AND SOFT DATA

Megan Lynne Liedtke Tesar, Ph.D.

University of Nebraska, 2011

Advisor: David B. Marx

The overall goal of this research, which is common to most spatial studies, is to predict a value of interest at an unsampled location based on measured values at nearby sampled locations. To accomplish this goal, ordinary kriging can be used to obtain the best linear unbiased predictor. However, there is often a large amount of variability surrounding the measurements of environmental variables, and traditional prediction methods, such as ordinary kriging, do not account for an attribute with more than one level of uncertainty. This dissertation addresses this limitation by introducing a new methodology called weighted kriging. This prediction technique accounts for measurements with significant variability, i.e., soft data, in addition to measurements with little or no variability, i.e., hard data.

To investigate the differences between weighted kriging and ordinary kriging, a simulation study was conducted. Validation statistics were used to evaluate and compare the prediction procedures, and it was found that weighted kriging yields more desirable

results than traditional kriging methods. As a follow-up, the prediction procedures were compared using real data from a groundwater quality study.

Bayesian Maximum Entropy (BME) is then introduced as an alternative method to utilize soft data in prediction. Numerical implementation of this approach is possible with the Spatiotemporal Epistemic Knowledge Synthesis-Graphical User Interface (SEKS-GUI). Using this interface, two simulation studies were conducted to investigate the differences between BME and weighted kriging. In the first study, probabilistic soft data in the form of the Gaussian distribution were used. However, since proponents of the BME approach claim that it performs extremely well when the soft data are skewed, the second study used nonsymmetrical soft data generated using a triangular distribution. In both studies, the weighted kriging validation statistics were more desirable than those from BME.

ACKNOWLEDGEMENTS

This dissertation could not have been written without the help of my advisor, Dr. David B. Marx. I am grateful for the guidance and encouragement he offered throughout graduate school and thankful for the lessons he taught me along the way. I would also like to thank my other committee members, Dr. Stephen Kachman, Dr. Erin Blankenship, Dr. Roy Spalding, and Dr. Jeff Pedersen for their guidance and support throughout my academic program. I am also grateful to Dr. Alexander Kolovos, Dr. Mary Exner Spalding, and Steve Westerholt for their time, assistance, and invaluable suggestions.

On a more personal level, I would like to thank God, my family, and my friends. I would especially like to thank my husband and my parents for their patience, support, and encouragement during this time.

This research was supported by the United States Department of Agriculture Cooperative State Research, Education, and Extension Service (USDA-CSREES) Integrated Research, Education, and Extension Competitive Grants Program - National Integrated Water Quality Program Conservation Effects Assessment Project (CEAP), under Agreement No. 2006-51130-03708.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1 Introduction

The conservation and preservation of the world's natural resources are important issues in today's society. To properly address these issues and protect the environment, it is vital to accurately model and predict the environment's natural processes. According to Serre (1999), "past experience shows that measures taken to control water and air pollution have resulted in socio-economic benefits that far outweighed their cost, such as reduction of medical and remediation expenses and improved life conditions" (p. 1). However, accuracy is not always easy to achieve because there is often a large amount of variability surrounding the measurements of environmental variables, e.g., crop yield, groundwater nitrate levels, and precipitation (Olea, 2006). This variability leads to uncertain predictions, and consequently to uninformed decision making. It is therefore important to develop tools which account for measurements with significant variability, i.e., soft data, in addition to measurements with little or no variability, i.e., hard data. Traditional methods, such as ordinary kriging, however, do not account for an attribute with more than one level of uncertainty.

This dissertation consists of three chapters addressing the utilization of both hard and soft data in spatial prediction. As previously described, hard data are exact measurements or measurements with little or no variability. For example, suppose the rainfall is recorded in Lincoln, Nebraska on March 28, 2011, at 5:00 pm. Thunderstorms are moving through the area, and there is a rain gauge at three different locations in town.

The measurements are recorded and given by X_{hard} = (0.23, 0.58, 0.64), where P(x_{hard} = X_{hard}) = 1 (Serre, 2007).

Soft data, on the other hand, are measurements that include a significant amount of variability. This data may be of two types. The first is interval soft data, which are intervals with a lower bound **a** and upper bound **b** on the measurements (Serre, 2007). For example, at two data points, the concentration of a particular air matter is below the detection limit of 5 ppm. Thus, the soft data are given by **a** = (0, 0), **b** = (5, 5), and x_{soft} = (x_1, x_2) where P(**a** < x_{soft} < **b**) =1 (Serre, 2007). The second type is probabilistic soft data, which have uncertainty that can be described by a probability density function (pdf), a function which specifies the possible values of a random variable and their associated probabilities (Serre, 2007). Soft data may be due to measurement error, prediction error of the physical model, a secondary variable, mixing of data observed at different spatial/temporal scales, statistical estimates, or environmental sensors (Serre, 2007). Incorporating this uncertain data into the estimation and prediction process is important for accurate results, especially when the number of hard data points is limited.

1.1 Incorporating Soft Data in the Kriging Equations

Chapter 2 describes the overall goal of this research. The goal, which is common to most spatial studies, is to predict an attribute of interest at an unsampled location based on measured values at nearby sampled locations (Cressie, 1991). To accomplish this goal, three widely used semivariogram models are defined which can be used to model the spatial structure. For prediction, the ordinary kriging equations are defined, but their

inability to deal with soft data resulted in the derivation of weighted kriging equations. These equations incorporate soft data in the prediction procedure and are referred to as the weighted kriging equations because observations with different variability are weighted differently in the estimation of the semivariogram parameters.

A simulation study was used to investigate the differences between three prediction procedures. The first procedure used the ordinary kriging equations and only the hard data, the second procedure used the ordinary kriging equations and both the hard and soft data but treated them both as hard data, and the third procedure used the weighted kriging equations. Thus, the third procedure was the only one that incorporated the soft data and weighted it differently than the hard data. As a follow-up to the simulation, two of the three aforementioned prediction procedures, including the one which uses only the hard data and the one which uses both the hard and soft data in weighted kriging, were compared using real data from a groundwater quality study.

1.2 Weighted Kriging vs. Bayesian Maximum Entropy: Gaussian

Chapter 3 introduces the Bayesian Maximum Entropy (BME) approach. This approach also utilizes soft data in prediction and can be implemented with the Spatiotemporal Epistemic Knowledge Synthesis-Graphical User Interface (SEKS-GUI). A simulation study was conducted to investigate the differences between the BME approach and the weighted kriging equations. In this study, probabilistic soft data in the form of the Gaussian distribution were used.

1.3 Weighted Kriging vs. Bayesian Maximum Entropy: Triangular

Proponents of Bayesian Maximum Entropy claim that it performs extremely well when the soft data are skewed. Since symmetric soft data of the Gaussian form were used in Chapter 3, nonsymmetric soft data were generated in Chapter 4 to examine the aforementioned claim. To create skewed data, the soft data were generated using a nonsymmetrical triangular distribution. A final simulation study was conducted to compare the results obtained from BME to those produced from the weighted kriging equations.

1.4 References

Cressie, N. (1991). *Statistics for spatial data*. New York: John Wiley & Sons.

Olea, R. A. (2006). A six-step practical approach to semivariogram modeling. *Stochastic Environmental Research Risk Assessment, 20*, 307-318.

Serre, M.L. (1999). Environmental spatiotemporal mapping and ground water flow modelling using the BME and ST methods (Ph.D. Dissertation, University of North Carolina at Chapel Hill, 1999).

Serre, M.L. (2007, July). *Introduction to Bayesian maximum entropy*. Paper presented at the BME workshop sponsored by the Department of Statistics, University of Nebraska-Lincoln.

Chapter 2 Incorporating Soft Data into the Kriging Equations

2.1 Introduction

Spatial Statistics is the study of observations that are spatially located, that is each observation has a value for the attribute of interest as well as its spatial coordinates (Cressie, 1991). For example, a data set from a soil nutrient test may consist of the amount of iron, copper, and zinc in each soil sample along with the exact sampled location (longitude and latitude). According to Schabenberger and Gotway (2005), "the foremost reason for studying spatial statistics is that we are often not only interested in answering the "how much" question, but the "how much is where" question" (p. 1). Soil science, however, is not the only discipline to which spatial statistics is applicable. It is impossible to list all the disciplines that work with data collected from different spatial locations, but a few of them include geology, epidemiology, crop science, ecology, and astronomy (Cressie, 1991).

Often the goal in the study of these spatially correlated observations is to predict the value for the attribute of interest at an unsampled location based on measured values at nearby sampled locations (Cressie, 1991). In order to do this, three assumptions are necessary. The first assumption is that the sampled values are measured precisely and accurately (Clark & Harper, 2000). The second assumption is that the unsampled locations are part of a physically continuous and homogeneous surface, and the third assumption is that the values at the unsampled locations are related to one another in a

way which depends on the distance and direction between their locations (Clark & Harper, 2000).

If these assumptions are valid and there is a relationship between the values which depends on the location of the samples, a predicted value is produced which is superior to the arithmetic mean (Clark & Harper, 2000). This value relies on the fact that the unknown value is more strongly related to sample values which are close to it in terms of location (Cressie, 1991; Schabenberger & Gotway, 2005). In other words, low values are likely to be near other low values and high values are likely to be near other high values. The predicted value for an unknown value, Y, is constructed as a linear combination of the neighboring sample values. The simplest is a weighted average, where a sample of the closest neighboring observations are selected and combined with weighting factors. This weighted average of Y is be given by

$$Y^* = \sum_{i=1}^{n} w_i y_i \tag{2.1}$$

where n is the number of observations included in the sample, y_i are the values of the observations, and w_i are the weights given to each observation and are chosen according to how close each observation is to the unsampled location and their location to each other (Clark & Harper, 2000). Additionally,

$$\sum_{i=1}^{n} w_i = 1 \tag{2.2}$$

to ensure that the predictor is unbiased (Clark & Harper, 2000).

2.2 Semivariogram Models

A semivariogram is a function describing the relationship between sample values and the distance and possibly the direction between their locations. More specifically, it is defined to be half the expected squared difference between random variables separated by a specific distance and in a certain direction (Journel & Huijbregts, 1978). Several realizations of the random variables are necessary to estimate the semivariogram, but generally, only one realization is available. Thus, it is assumed that the semivariogram depends only on the separation vector h and not on the location of the points (Journel & Huijbregts, 1978). This assumption is called the intrinsic hypothesis or the hypothesis of second-order stationarity of the differences (Cressie, 1991; Journel & Huijbregts, 1978; Lee & Ellis, 1997). The result of this assumption is that, within the spatial domain, the structure of the variability between points separated by a distance smaller than the range is constant and independent of location (Journel & Huijbregts, 1978). By making this assumption, it is possible to estimate the semivariogram from the data with the following function:

$$\gamma(h) = \frac{1}{2N_h} \sum_{h} (y_i - y_j)^2 \tag{2.3}$$

where h denotes a specified distance and direction (Journel & Huijbregts, 1978). For each h, find all possible pairs of samples, denoted by N_h, and repeat the calculation for as many values of h as the sample will support.

In order to visualize the relationship, a semivariogram graph can be plotted for each direction. The semi-variances are plotted on the vertical axis with the distances between the samples on the horizontal axis. In the simplest situation, the semivariogram

is the same in any direction, and the spatial structure is called isotropic (Journel & Huijbregts, 1978; Olea, 2006). In this case, all experimental semivariograms are averaged, and the result is a semivariogram that is smoother than the individual directional semivariograms (Olea, 2006). However, if the semivariogram has directional properties, then anisotropy is present (Cressie, 1991). As a result, the semivariogram model accounts for the varying spatial structure by direction. One such study that may be anistropic in nature is a pollution study where flow directions must be taken into consideration (Clark & Harper, 2000). The two major types of anisotropy are geometric and zonal. In geometric anisotropy, the range differs by direction, whereas in zonal anisotropy, the sill differs by direction (Journel & Huijbregts, 1978). It is also possible to have a mixture of geometric and zonal anisotropy.

Many models exist to describe the theoretical semivariogram, but a special class of functions must be considered for the kriging minimization problem. If the coefficient matrix is not singular, any quadratic minimization problem has a unique, positive solution (Olea, 2006). Thus, for kriging, the semivariogram must be negative definite (Olea, 2006). A negative definite model prevents singular kriging matrices and negative prediction variances. There is no guarantee that a sample semivariogram satisfies this property, but the basic shape of the semivariogram limits the functions of interest (Olea, 2006). Here these include the Spherical, Exponential, and Gaussian models.

The Spherical model is cubic and relies on two parameters, the range and the sill. There may also be a third parameter, the nugget effect, a positive intercept on the vertical axis (Clark & Harper, 2000). The equation for this model is given by:

$$\gamma(h) = \begin{cases} C_0 + C\left\{1.5\frac{h}{a} - 0.5\left(\frac{h}{a}\right)^3\right\} & \text{when } 0 < h \le a \\ C_0 + C & \text{when } h > a \end{cases} \tag{2.4}$$

where γ is the semivariogram value, h is the distance between two points, a is the range, C_0 is the nugget effect, C is the partial sill, and $C_0 + C$ is the sill of the spherical component (Clark & Harper, 2000). The nugget can be described as the vertical jump from the value of 0 at the origin to the value of the semivariogram at extremely small separation distances and is due to measurement error and variability of the sampled property (Journel & Huijbregts, 1978). The sill is the plateau the semivariogram reaches at the range, and the range is the distance at which the semivariogram stops increasing (Clark & Harper, 2000). Furthermore, observations separated by a distance greater than the range can be assumed to be spatially independent. As shown in Figure 2.1, the upper asymptote of the model is $C_0 + C$ (Clark & Harper, 2000).

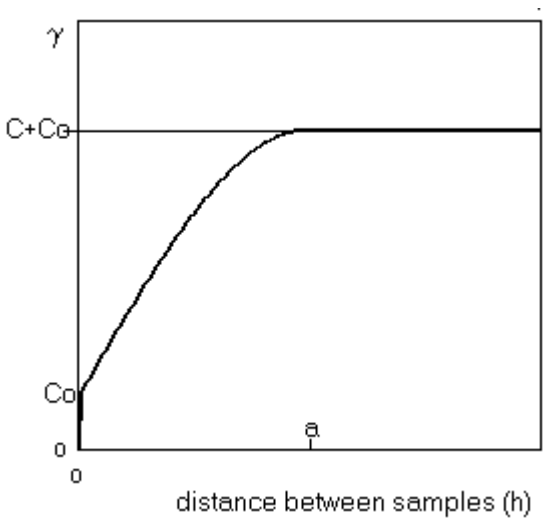

Figure 2.1: Spherical semivariogram

The Exponential model, like the Spherical model, relies on two main parameters, the range and the sill. In addition, there may also be a nugget effect. The equation for this model is given by:

$$\gamma(h) = C_0 + C\left\{1 - \exp\left(\frac{-h}{a}\right)\right\} \text{ when } h > 0 \qquad (2.5)$$

where γ is the semivariogram value, h is the distance between two points, a is the range, C_0 is the nugget effect, C is the partial sill, and $C_0 + C$ is the sill of the exponential component (Clark & Harper, 2000). The range, however, does not represent the distance at which observations become independent. Instead, the Exponential model reaches about two-thirds of its height at a distance of a and must go three times this distance to reach its asymptotic sill (Journel & Huijbregts, 1978). The shape for this model is shown in Figure 2.2 (Clark & Harper, 2000).

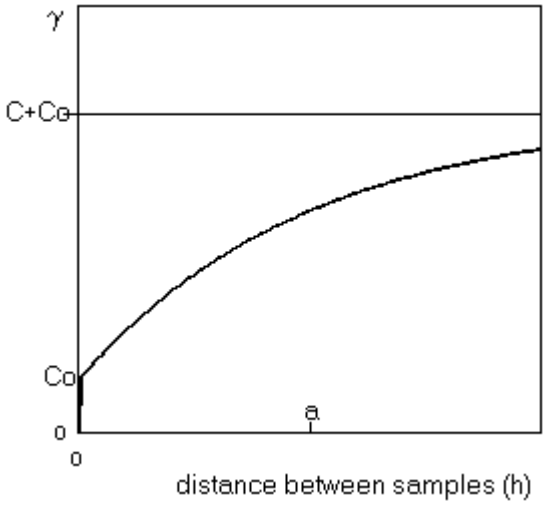

Figure 2.2: Exponential semivariogram

The Gaussian model is commonly used to represent events with a small scale spatial structure (Clark & Harper, 2000). The equation for this model is similar to the Normal cumulative distribution function and given by:

$$\gamma(h) = C_0 + C\left\{1 - \exp\left(\frac{-h^2}{a^2}\right)\right\} \text{ when } h > 0 \tag{2.6}$$

where γ is the semivariogram value, h is the distance between two points, a is the range, C_0 is the nugget effect, C is the partial sill, and $C_0 + C$ is the sill of the Gaussian component (Clark & Harper, 2000). Again, the range does not represent the distance at which observations become independent. According to Journel and Huijbregts (1978), the Gaussian model reaches about two-thirds of its height at a distance of a and reaches its asymptotic sill at a distance of $a\sqrt{3}$. The shape for this model is shown in Figure 2.3 (Clark & Harper, 2000).

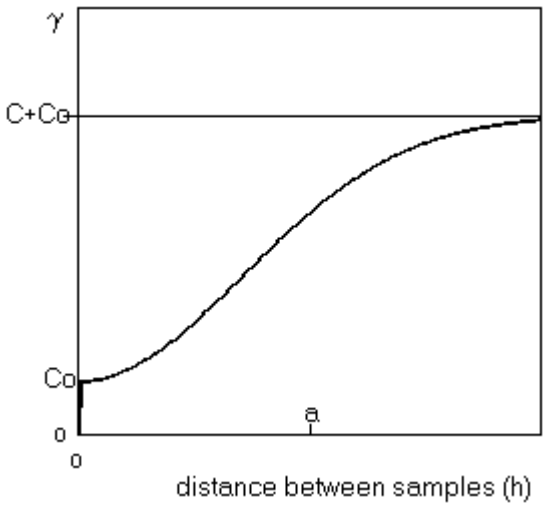

Figure 2.3: Gaussian semivariogram

For all three models, the value of the semivariogram for a distance of zero is zero. However, due to sampling error and scale variability the values recorded at extremely small separation distances may be rather dissimilar causing discontinuity at the origin (Clark & Harper, 2000). As mentioned previously, this vertical jump from zero to these values is referred to as "the nugget effect" (Isaaks & Srivastava, 1989), and must also be considered during spatial analyses.

Although the Spherical, Exponential, and Gaussian models are defined in terms of their semivariogram functions, it is also possible to model them using covariance functions. If it is assumed that the variables in our random function model have the same mean and variance, the following relationship exists between the semivariogram and the covariance:

$$\gamma_{ij}(h) = \tilde{\sigma}^2 - \tilde{C}_{ij}(h) \tag{2.7}$$

where $\gamma_{ij}(h)$ is the semivariogram value between points i and j separated by a distance and direction h, $\tilde{\sigma}^2$ is the sill or the variance of the random function model, and $\tilde{C}_{ij}(h)$ is the covariance between points i and j separated by a distance and direction h (Clark & Harper, 2000). Figure 2.4 compares the Spherical, Exponential, and Gaussian covariance functions.

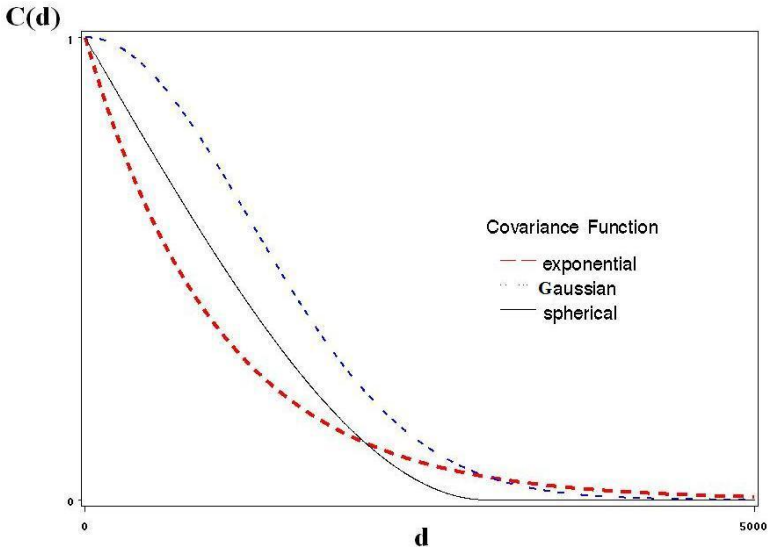

Figure 2.4: Comparison of covariance functions

Since the Spherical covariance function is most common in agricultural studies, the examples provided in this paper use the Spherical semivariogram function and assume no nugget effect. In addition, the example datasets are simulated under isotropic conditions, i.e., under the assumption that the spatial variability is the same in all directions.

2.3 Ordinary Kriging

When predicting an unsampled location, the goal is to produce a weighted average from neighboring samples. To calculate these weights, the method of ordinary kriging is used. The theory behind kriging was developed in 1963 in a work entitled "Principles of Geostatistics" by George Matheron, a French mathematician who became known as the founder of spatial statistics. This method is a local prediction technique which provides the best linear unbiased predictor (Journel & Huijbregts, 1978). In other

words, it aims to minimize the variance of the errors, the predicted values are weighted linear combinations of the data, and the difference between the predictor's expected value and the true value of the attribute being predicted is equal to zero (Journel & Huijbregts, 1978). However, the error variance and the mean residual are unknown so a probability model is used to calculate the error variance when the bias is zero. In our case, the Spherical semivariogram model will be used, but other negative definite models can easily be used in its place. The weights are then chosen for nearby samples to ensure that the modeled error variance is minimized and the average error for the model is zero.

The ordinary kriging equations to predict the attribute of interest at an unsampled location, Y, in matrix form are $\mathbf{C} \cdot \mathbf{w} = \mathbf{D}$, where **C**, **w**, and **D** are defined by Clark and Harper (2000) as follows:

$$\begin{bmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1n} & 1 \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2n} & 1 \\ \vdots & & \ddots & & \vdots \\ \gamma_{n1} & \gamma_{n2} & \cdots & \gamma_{nn} & 1 \\ 1 & 1 & \cdots & 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} w_1 \\ \vdots \\ w_n \\ \mu \end{bmatrix} = \begin{bmatrix} \gamma_{10} \\ \vdots \\ \gamma_{n0} \\ 1 \end{bmatrix} \tag{2.8}$$

The semivariogram values are denoted by γ_{ij}, and n is the number of nearest neighbors used in prediction. Thus, **C** consists of the semivariogram values between all pairs of observations used in prediction, where the last row and column of **C** provide a linear constraint that the weights sum to one and ensure unbiasedness. Vector **w** consists of the weights and μ, the Lagrange parameter, and matrix **D** consists of the semivariogram values between each observation and the unsampled location. To solve for the weights, multiply both sides by $\mathbf{C}^{-1}$ to produce the solution vector $\mathbf{w} = \mathbf{C}^{-1}\mathbf{D}$. Since only known

negative definite functions are considered to fit the semivariogram, the existence of $\mathbf{C}^{-1}$ is guaranteed. A generalized inverse may have to be used if the distances between the observations used in prediction and the point to be predicted are larger than the range. However, observations farther than the range are generally not used in prediction. The predicted value is then expressed as $Y^* = \mathbf{w}'\mathbf{y}$, where $\mathbf{y}$ represents the vector of sample values used in prediction with a zero in the $(n+1)$ position, and the prediction variance is given by $\sigma^2 = \mathbf{w'D} - \gamma(0)$ (Clark and Harper, 2000).

2.4 Ordinary Kriging Limitations and Alternative Methods

Although ordinary kriging provides the best linear unbiased predictor, it does have its limitations. In particular, kriging does not provide a framework to incorporate data of differing precisions. "As a consequence, these methods lack the theoretical underpinnings and practical flexibility to account for important sources of physical knowledge" (Serre, 1999, pg. 2). Furthermore, kriging assumes the data are Gaussian, and although it is the best among linear predictors, it is not necessarily the best when compared to non-linear predictors (Serre, 2007).

Nonlinear prediction techniques, including disjunctive kriging and multivariate Gaussian kriging, were proposed by Journel and Huijbregts in 1978. These methods require Gaussian-related hypotheses so Matheron and Christakos developed more general predictors in the 1980's (Christakos, 1990). However, none of these methods incorporate prior information into the analysis. (Christakos, 1990)

To account for the uncertain information, soft kriging was proposed by Journel (1986). This method "allows a coding of both hard data and constraint intervals as prior cumulative distribution functions (cdfs) which are interpolated into posterior cdfs" (Journel, 1986, p. 269). The interpolation procedure is done by means of least squares and leads to the derivation of non-Gaussian confidence intervals and estimates of posterior probability distributions (Journel, 1986). However, this approach lacks a firm rule for assigning prior probabilities, requires a large amount of statistical inference, and assumes Gaussian probability distributions (Christakos & Li, 1998). Furthermore, a significant amount of information may be lost due to approximations (Christakos & Li, 1998).

2.5 Simulation Study

The purpose of this study was to overcome the limitations of the aforementioned prediction procedures. In order to accomplish this, a weighted kriging procedure which utilizes both hard and soft data was derived. This prediction procedure was then compared to two ordinary kriging methods. The first kriging method used only hard data, and the second procedure used both the hard and soft data but treated both as hard. The simulation study was conducted in SAS® Version 9.2 (SAS Institute, 2008) to investigate these methods and the potential differences between them.

To begin the simulation study, a Spherical spatial floor was generated on a 40 by 40 grid (1600 points). In all simulations, the Spherical spatial structure had a sill of 1.0 and a nugget of 0. These simulation parameters were chosen to ensure a strong spatial

structure. As shown in Table 2.1, the range was either 15 or 30. From the 1600 points, 400 were randomly selected to be observed values and the remaining 1200 were used for validation. Either 10% or 50% of the observed values were randomly chosen to be soft data. Thus, the hard data consisted of 360 observations when 10% of the data were soft and 200 observations when 50% of the data were soft. The remaining 40 or 200 observations, respectively, made up the soft data. These points became soft by adding an independent Normal component with a mean of 0 and variance of 0.5. For each combination in Table 2.1, 105 data sets were simulated.

Nugget	**Sill**	**Range**	**% Soft Data**
0	1	15	10
0	1	15	50
0	1	30	10
0	1	30	50

Table 2.1: Simulation parameters used to compare different ranges and different percentages of soft data

The semivariogram of each combination listed above was estimated in three different ways. The first procedure modeled the Spherical semivariogram based only on the hard data, the second used both the hard and soft data but treated both as hard, and the third procedure used both the hard and soft data but weighted the observations in the semivariogram estimation based on the type of data (hard or soft).

2.5.1 Hard Data

First consider the procedure which used only the hard data or only those observations to which no additional variability was added. Using these observations, SAS® PROC VARIOGRAM was implemented to find all possible pairs of hard data points. Then nonlinear least squares was used to model the following equations for the Spherical semivariogram:

$$\gamma(h) = \begin{cases} C_0 + C \cdot Q & \text{when } 0 < h \le a \\ C_0 + C & \text{when } h > a \end{cases} \tag{2.9}$$

where

$$Q = \left\{ 1.5\frac{h}{a} - 0.5\left(\frac{h}{a}\right)^3 \right\} \text{ when } 0 < h \le a. \tag{2.10}$$

The next step was to check the quality of the range and sill parameters that resulted. This was done by classifying the results into one of three categories. The first category indicated that the procedure converged correctly, the second indicated incorrect convergence, and the third indicated that the procedure failed to converge. In general, incorrect convergence meant that the range and/or sill parameters were outside specified limits. In particular, for the data sets with a range of 15, incorrect convergence meant that the range was less than 1 or greater than 50 and/or the sill was less than 0.1 or greater than 20. For the data sets with a range of 30, incorrect convergence meant the range was less than 1 or greater than 100 and/or the sill was less than 0.1 or greater than 20. If the results fell into the second or third category and only the hard data was being used, the parameters were adjusted as follows. The nugget was changed to 0, the sill to 1, and the

range to 15 or 30. In other words, the parameters were changed to correspond to the values used in the simulation.

The final step of this procedure was to apply PROC KRIGE2D to perform ordinary kriging. The 20 nearest neighbors of the hard data points were used to produce kriging predictions at all 1600 points on the 40 by 40 grid. The kriging predictions and corresponding standard errors were then used to compute the validation statistics given in Section 6 of Chapter 2.

2.5.2 Hard and Soft Data Treated as Hard

The second procedure used both the hard and the soft data but treated all observations as hard. Thus, this procedure used all 400 of the observed values in PROC VARIOGRAM. Then nonlinear least squares was used to model the following equations for the Spherical semivariogram:

$$\gamma(h) = \begin{cases} C_0 + C \cdot Q & \text{when } 0 < h \le a \\ C_0 + C & \text{when } h > a \end{cases} \tag{2.11}$$

where

$$Q = \left\{ 1.5\frac{h}{a} - 0.5\left(\frac{h}{a}\right)^3 \right\} \text{ when } 0 < h \le a. \tag{2.12}$$

The quality of the parameters was checked, and if they were outside the specified limits mentioned in 2.5.1 or if the procedure failed to converge, the nugget was changed to 0, the sill to 1, and the range to 15 or 30. Again, PROC KRIGE2D was used to perform ordinary kriging. The 20 nearest neighbors of the 400 observed values were

used to produce kriging predictions at all1600 points on the 40 by 40 grid. The kriging predictions and corresponding standard errors were then used to compute the validation statistics given in Section 6 of Chapter 2.

2.5.3 Weighted Kriging

The third procedure used both the hard and soft data but weighted the observations in the semivariogram estimation. Again, PROC VARIOGRAM was used to find all possible pairs of the 400 observed values. However, unlike the previous two situations, this procedure required the use of an additional variable to distinguish whether or not the pairs consisted of two hard data points, one hard and one soft, or two soft data points. The type of data in each pair indicated which equation to use in iteratively reweighted least squares to model the Spherical semivariogram.

To develop these equations, first consider the following notation. Let a hard data point at location k be denoted by x_k^H and a soft data point at location k be denoted by x_k^S. Next, let the variance of a hard data point be denoted by $V(x_k^H)$ and the variance of a soft data point be denoted by $V(x_k^S)$. Finally, let the covariance between two points separated by a distance of h be denoted by $\mathrm{Cov}(x_k, x_{k+h})$. Then define the quantities as follows:

$$V(x_k^H) = \sigma^2,$$

$$V(x_k^S) = \sigma^2 + \Delta,$$

$$Cov(x_k, x_{k+h}) = C(h).$$

Thus, the semivariogram value between two hard data points is given by

$$\begin{aligned}\frac{1}{2}\left[V(x_k^H - x_{k+h}^H)\right] &= \frac{1}{2}\left[V(x_k^H) + V(x_{k+h}^H) - 2Cov(x_k^H, x_{k+h}^H)\right] \\ &= \frac{1}{2}\left[\sigma^2 + \sigma^2 - 2C(h)\right] \\ &= \sigma^2 - C(h) \\ &= \gamma_{HH}(h).\end{aligned} \tag{2.13}$$

Whereas, the semivariogram value between one hard observation and one soft observation is given by

$$\begin{aligned}\frac{1}{2}\left[V(x_k^H - x_{k+h}^S)\right] &= \frac{1}{2}\left[V(x_k^H) + V(x_{k+h}^S) - 2Cov(x_k^H, x_{k+h}^S)\right] \\ &= \frac{1}{2}\left[\sigma^2 + (\sigma^2 + \Delta) - 2C(h)\right] \\ &= \sigma^2 + \frac{1}{2}\Delta - C(h) \\ &= \gamma_{HS}(h),\end{aligned} \tag{2.14}$$

and the semivariogram value between two soft data points is

$$\begin{aligned}\frac{1}{2}\left[V(x_k^S - x_{k+h}^S)\right] &= \frac{1}{2}\left[V(x_k^S) + V(x_{k+h}^S) - 2Cov(x_k^S, x_{k+h}^S)\right] \\ &= \frac{1}{2}\left[(\sigma^2 + \Delta) + (\sigma^2 + \Delta) - 2C(h)\right] \\ &= \sigma^2 + \Delta - C(h) \\ &= \gamma_{SS}(h).\end{aligned} \tag{2.15}$$

In summary, the adjusted semivariogram values are as follows:

$$\begin{aligned}\gamma_{HH}(h) &= \gamma(h), \\ \gamma_{HS}(h) &= \gamma(h) + \frac{1}{2}\Delta, \\ \gamma_{SS}(h) &= \gamma(h) + \Delta.\end{aligned} \tag{2.16}$$

Now, let Δ = nugget. Thus,

$$\begin{aligned} \gamma_{HH}(h) &= \gamma(h), \\ \gamma_{HS}(h) &= \gamma(h) + \frac{1}{2} nugget, \\ \gamma_{SS}(h) &= \gamma(h) + nugget. \end{aligned} \tag{2.17}$$

Iteratively reweighted least squares was then used to estimate the Spherical semivariogram model given in equation (2.4). The points used for this estimation consisted of all pairs of the observed values. The distance between each pair of observed values served as the independent variable and their squared difference in attribute value served as the dependent variable. Since each pair consisted of two hard observations, one hard and one soft observation, or two soft observations, differing weights were assigned to each pair based upon which of these three conditions was true. More specifically, the weights were based on the semivariogram equations defined in equation (2.17) and equal to the reciprocal of the square root of the variance. Thus, if there were two hard data points in the pair, the weight was the reciprocal of the square root of the variance of independent hard data observations, i.e. the sill = σ^2:

$$\text{weight}(x_k^H,\ x_{k+h}^H) = \frac{1}{\sqrt{\sigma^2}}. \tag{2.18}$$

If the pair consisted of one hard and one soft observation, the weight was defined as

$$\text{weight}(x_k^H,\ x_{k+h}^S) = \frac{1}{\sqrt{\sigma^2 + \frac{1}{2} \cdot \text{nugget}}}, \tag{2.19}$$

and if the pair consisted of two soft observations, the weight was defined as

$$\text{weight}(x_k^S,\ x_{k+h}^S) = \frac{1}{\sqrt{\sigma^2 + \text{nugget}}}. \tag{2.20}$$

Following estimation of the semivariogram, the quality of the parameters was checked. If they were outside the specified limits or if the procedure failed to converge, the nugget was changed to 0.5, the sill to 1.25, and the range to 15 or 30. In the previous two procedures the nugget was changed to 0 and the sill to 1, but to account for the presence of soft data, the nugget and sill were increased based on how the soft data was constructed, i.e., by adding an independent N(0, 0.5) component.

Again, PROC KRIGE2D was used to find the 20 nearest neighbors of the observed values. Then a loop was used to predict all 1600 points. Within the loop, the 20 nearest neighbors of each point were used by PROC IML to form matrix **C** and matrix **D** in equation (2.8). These matrices were different from those constructed in the previous two procedures in that they relied on the semivariogram values as defined in equation (2.17). Thus, the semivariogram values in matrix **C** were calculated based on whether the value corresponded to a pair of hard data points, a pair of soft data points, or one hard and one soft data point. Likewise, the semivariogram values in matrix **D** were based upon whether each particular observation used to predict the unsampled location was hard or soft. These matrices were used to produce the solution vector $\mathbf{w} = \mathbf{C}^{-1}\mathbf{D}$. Since the Spherical model was used and the observations included in prediction, the 20 nearest points, were within the range of the unsampled location, the existence of $\mathbf{C}^{-1}$ was

guaranteed. Then the predictions, Y* = **w'y**, and their corresponding standard errors were calculated and used to form the validation statistics given Section 2.6.

2.6 Results

Prior to running the simulation, the expectations were that the weighting kriging procedure would perform the best. In other words, it would result in the most desirable validation statistics. In addition, the difference between the methods was expected to be larger when 50% of the data are soft rather than only 10%.

To compare the three procedures, the following validation statistics were used:

$$\text{Root Mean Square Error (RMSE)} = \sqrt{\frac{\sum\left[\text{Actual-Predicted}\right]^2}{n}} \tag{2.21}$$

$$\text{Average Variance (AVAR)} = \frac{\sum \sigma_{p_i}^2}{n} \text{ where } \sigma_{p_i}^2 \text{ is the prediction variance of point } i \tag{2.22}$$

$$\text{Standardized Mean Error (SME)} = \frac{\sum\ [(\text{Actual-Predicted}) / \sigma_{p_i}]}{n} \tag{2.23}$$

$$\text{Absolute Mean Prediction Error (ABSMPE)} = \frac{\sum\left|\text{Actual-Predicted}\right|}{n} \tag{2.24}$$

$$\text{Root Mean Square Standardized Error (RMSSE)} = \sqrt{\frac{\sum\left[\ (\text{Actual-Predicted}) / \sigma_{p_i}\ \right]^2}{n}} \tag{2.25}$$

In each equation, the predicted quantities refer to the results obtained from the kriging procedures described in Sections 2.5.1-2.5.3. The RMSE and the AVAR, the average of all the prediction variances, should be small for a model which fits the data well. The SME is the only fit statistic that can be negative but should be close to zero for a good fitting model. Furthermore, the ABSMPE should be close to zero while the RMSSE should be close to one. If the RMSSE is large, the variability in our predictions is underestimated, but if it is less than one, then this variability is overestimated.

Tables 2.2-2.5 summarize the means of the validation statistics from the simulation study. **Hard** indicates that only the hard data were used to obtain the semivariogram estimates and to predict unobserved values. **Both** indicates that both the hard and soft data were treated as hard data in estimation and prediction, and **Weighted** indicates the use of the weighted kriging procedure. Friedman's Chi-Square Test was used to determine if there was a significant difference between the three prediction procedures. For this nonparametric test, each simulated data set served as a block, and the prediction technique was the treatment.

	RMSE	AVAR	SME	ABSMPE	RMSSE
Hard	**10.2109**	**0.1699**	-0.0100	**1.0862**	**3.1293**
Both	**7.6252**	0.2236	-0.0095	**0.7792**	**1.7980**
Weighted	0.3737	0.2567	-0.00005	0.2953	0.7585

Table 2.2: Fit statistics obtained from ordinary kriging with hard data, ordinary kriging with hard and soft data treated as hard, and weighted kriging with range=15 and 10% soft data
Note: Bold indicates significantly different than Weighted at alpha level of 0.05

	RMSE	AVAR	SME	ABSMPE	RMSSE
Hard	**12.6805**	**0.2373**	0.0182	**1.2806**	**2.8038**
Both	**0.4431**	**0.4400**	-0.0015	**0.3514**	**0.6920**
Weighted	0.4563	0.3000	-0.0051	0.3267	0.8030

Table 2.3: Fit statistics obtained from ordinary kriging with hard data, ordinary kriging with hard and soft data treated as hard, and weighted kriging with range=15 and 50% soft data
Note: Bold indicates significantly different than Weighted at alpha level of 0.05

	RMSE	AVAR	SME	ABSMPE	RMSSE
Hard	**85.9100**	**0.1625**	-0.0042	**3.5719**	**4.0665**
Both	**23.7644**	0.1576	0.0172	**1.7193**	**4.3161**
Weighted	0.2919	0.1333	-0.00004	0.2130	0.7764

Table 2.4: Fit statistics obtained from ordinary kriging with hard data, ordinary kriging with hard and soft data treated as hard, and weighted kriging with range=30 and 10% soft data
Note: Bold indicates significantly different than Weighted at alpha level of 0.05

	RMSE	AVAR	SME	ABSMPE	RMSSE
Hard	**41.3732**	**0.1476**	**0.1081**	**3.1984**	**9.4365**
Both	**5.6448**	**0.2494**	-0.0095	**0.7806**	2.4385
Weighted	0.2918	0.1602	-0.0009	0.2305	0.7606

Table 2.5: Fit statistics obtained from ordinary kriging with hard data, ordinary kriging with hard and soft data treated as hard, and weighted kriging with range=30 and 50% soft data
Note: Bold indicates significantly different than Weighted at alpha level of 0.05

The issue of convergence was not a major concern in the simulation study. The data sets with a simulation range of 15 and 10% soft data converged correctly 95% of the time. Those with a range of 15 and 50% soft data converged correctly 92% of the time. The respective percentages were lower for the data sets with a simulation range of 30. When 10% of the data was soft, 84% of the data sets converged correctly, and when 50% of the data was soft, 81% converged correctly. Several of the simulations which did not converge were examined, and it was determined convergence would have been achieved if the number of iterations was increased from the default value of 100 or if a different set of starting values was defined. However, the change in the semivariogram model estimates was minimal when these changes were made. Furthermore, recall that the quality of the semivariogram estimates was checked, and the estimates were redefined if they were outside the specified limits. Thus, the Spherical semivariogram estimates which resulted after 100 iterations were utilized in this study.

2.7 Conclusions

In summary, the weighted kriging RMSE was significantly smaller than the other two procedures except when the range was 15 and 50% of the data were soft. In this case, ordinary kriging with both types of data resulted in the lowest RMSE, 0.4431, but the weighted kriging RMSE was only slightly larger at 0.4563. The weighted kriging SME was closest to zero except when the range was 15 and 50% of the data were soft. In this case, kriging with both types of data resulted in the SME closest to zero (-0.0015), but the weighted kriging SME (-0.0051) was not significantly different. In each

simulation, as desired, the weighted kriging ABSMPE was the smallest and the RMSSE was closest to one.

In three out of four cases, ordinary kriging with the hard data alone resulted in the smallest AVAR. However, this was to be expected as the slightly larger values produced by weighted kriging were caused by incorporating the more variable soft data in prediction. These higher prediction errors will be most evident in areas where only soft data contribute to the predictions and where hard data are limited. According to Kolovos (personal communication, February 7, 2008), this "informed" uncertainty is preferred over the "systematic" uncertainty which arises in ordinary kriging when predictions lie far away from any hard data. Thus, this uncertainty is preferred over the fictitiously lower prediction errors which resulted when the soft data were ignored in ordinary kriging.

In addition to comparing the three types of analyses, the four simulation cases with varying ranges and percentages of soft data were compared. The difference between the three procedures appeared to be most noticeable when 50% of the data were soft rather than 10%. Furthermore, the means resulting from the simulated data sets with a range of 15 were considerably smaller than those resulting from the data sets with a range of 30. Overall, weighted kriging performed the best.

2.8 Two-Step Kriging

If it were not possible to incorporate the soft data into the kriging equations, one may have proposed the following two-step approach. First, krige the unsampled

locations using only the hard data. This results in predicted values called $\hat{y}_{hard}$ with a prediction variance of $\hat{s}^2_{hard}$. Then, krige using the soft data and call these predicted values $\hat{y}_{soft}$ with a prediction variance of $\hat{s}^2_{soft}$. Each unsampled location now has two predictions, one soft and one hard. To obtain the predicted value for each point, a weighted average of the two would be used. The weights would be derived by minimizing the variance such that the weights sum to 1. Thus, the resulting predicted value at a particular unsampled location would be given by:

$$\hat{y} = \frac{\hat{s}^2_{soft}}{\hat{s}^2_{hard} + \hat{s}^2_{soft}} \hat{y}_{hard} + \frac{\hat{s}^2_{hard}}{\hat{s}^2_{hard} + \hat{s}^2_{soft}} \hat{y}_{soft}. \tag{2.26}$$

Although this approach may seem reasonable, the predicted value at the unsampled location would not only be vastly different, but it would also be less accurate than the result obtained from weighted kriging. To see why this is true, consider Figure 2.5. The location to be predicted is denoted by the letter P, while the letter S represents soft data points and H represents hard data points.

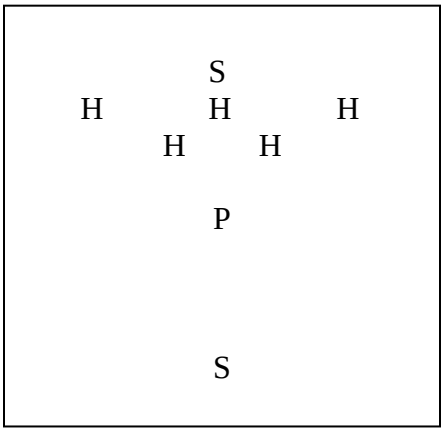

Figure 2.5: Hypothetical data plot

If the hard and soft data points were kriged separately as described above, the two soft points would have the same weight because they are the same distance from the unsampled location. However, if the hard and soft data points were combined and all of the points were considered for weighting, then the soft data point above the unsampled location would have a much smaller weight due to the fact that it is blocked by hard data. In addition, there is only one observed value below the unsampled location so that soft data point would have a larger weight as it provides valuable information that cannot be supplied by any other observed value. Thus, considering all points together in weighting kriging is better than kriging the hard and soft data separately.

2.9 Application to Groundwater Nitrate Concentrations

As a follow-up to the simulation study, the prediction techniques were compared using real data. The purpose of this application was to compare the results obtained from kriging with only hard data to the results obtained from using both hard and soft data in the weighted kriging procedure. The data used in this application came from a United States Department of Agriculture Cooperative State Research, Education, and Extension Service (USDA-CSREES) Conservation Effects Assessment Project (CEAP). The goal of this project was to assess agricultural conservation practices on groundwater quality by sampling the groundwater nitrate concentrations (mg/L) in Nebraska's central Platte River valley. Only those observations in the primary aquifer and in the northern section of the study area were considered in this application. In addition, the years of interest included 2003-2006.

To define the difference between hard and soft data in this application, consider the following scenario. Suppose the groundwater nitrate concentrations were measured at ten coordinates in 2005. In 2006, the concentrations were again measured at ten coordinates. However, four of the ten measurements in 2006 were at the exact same locations as measurements taken in 2005, leaving six locations that were measured in 2005 but not in 2006. These six measurements can be used as soft data to predict unknown values in 2006. Since they were not measured in 2006, there is more variability associated with these observations. However, they still provide valuable information regarding the nitrate concentrations at those locations and should not be ignored.

In this particular scenario, the addition of these non co-located 2005 observations as soft data to the 2006 data set increased the number of observations from 10 to 16. When this additional information is used to predict the nitrate concentration at an unsampled location, the standard error associated with the predicted value should be smaller. In other words, incorporating this data into the prediction process should lead to more precise results.

The increase in precision becomes increasingly apparent when hard data points are limited. For example, assume that ten hard data points exist within the range of an unsampled location. Based on these ten points, the resulting predicted value will be fairly precise. However, if there are only four hard data points within the range, the resulting value based on fewer observations will be much less precise. In fact, it is common procedure to require the use of at least 6-8 observations. Thus, if there are an additional four soft data points within the range of the unsampled location, eight data points (four

hard and four soft) are now available. Using all eight of these observed values yields a more accurate and precise prediction than using the four hard data points alone.

In the USDA-CSREES's CEAP study, the number of hard data points was much larger than in the scenario described above. In 2004, the groundwater nitrate concentrations were measured at 744 locations. In 2005, concentrations were measured at 671 locations, and in 2006, 625 locations were measured. A summary of these observations is given in Table 2.6.

Measured Year	Number of Observations	Mean	Standard Deviation	Minimum	Maximum
2004	744	21.96	8.33	0.10	47.00
2005	671	22.02	8.64	0.10	46.00
2006	625	21.29	8.57	0.10	47.10

Table 2.6: Summary of hard data nitrate concentrations (mg/L) from USDA-CSREES's CEAP study

As described above, the groundwater nitrate concentrations recorded in a previous year were used to predict unsampled locations in the current year. For example, 216 of the 671 observations in 2005 were recorded at locations that were not measured in 2006. Thus, these 216 observations were used as soft data for predicting the 2006 measurements. There was more variability associated with these observations since the measurements were recorded in the previous year. However, they provided valuable information regarding the nitrate concentrations at those locations. A summary of the soft data from the USDA-CSREES's CEAP study is provided in Table 2.7.

Measured Year	Number of Observations	Mean	Standard Deviation	Minimum	Maximum
2003	153	20.95	11.10	0.200	64.60
2004	178	21.00	7.26	1.40	35.80
2005	216	20.06	8.90	0.10	42.80

Table 2.7: Summary of soft data nitrate concentrations (mg/L) from USDA-CSREES's CEAP study

2.9.1 Methods

In order to improve the quality of the soft data points, an adjustment was made to take into consideration the change in nitrate concentrations over time. For example, the 2006 hard data points had an average value of 21.29 while the 2005 soft data points had an average value of 20.06. Thus, there was a difference of 1.23, and since the 2005 soft data points were used to predict 2006 values, the 2005 soft data points were adjusted by adding 1.23. Likewise, the 2004 soft data were adjusted by adding 1.02, and the 2003 soft data were adjusted by adding 1.01.

For the procedure using only hard data points, the closest 12 observations to the point to be predicted were considered. A variable was created to distinguish whether or not a point used in the prediction was a "quality" point. If the distance between the observed point and the unsampled location was less than or equal to the range, then the point was considered a "quality" point. However, if the distance was greater than the range, it was not a "quality" point. For an accurate prediction, at least 8 "quality" points were required for each predicted value.

When both the hard and soft data were considered, the closest 12 hard data points were again included in the prediction procedure. The distance between each of the 12 hard data points and the point to be predicted was calculated, and the maximum distance was used to determine which soft data points were included. First, the closest 12 soft data points were considered. Then, the distance between each of the 12 soft points and the point to be predicted was compared to the maximum distance described above. If the distance between the soft data point and the point to be predicted was larger than the maximum distance, then the soft data point was not used in predicting that point. Alternatively, if the distance between the points was less than or equal to the maximum distance, then the soft data was used in prediction. Therefore, anywhere from 0 to 12 additional soft data points were used in prediction, but at least 8 "quality" hard data points were required for each predicted value.

2.9.2 Results

Using the methodology described in 2.9.1, 523 quality values were predicted in the study area in 2004 and 2005. Due to the smaller number of hard observations, 519 quality values were predicted in 2006. A summary of the prediction results by year is provided in Tables 2.8-2.10. Each table contains four variables and their corresponding means. The first and second rows of each table contain the means of the predicted values. The mean in the first row was obtained from the procedure which used only the hard data (the data measured in the year to be predicted), and the mean in the second row was obtained from the procedure which used both the hard data and the soft data (the data

measured at different locations in the year prior to the year to be predicted). The third and fourth rows of each table contain the means of the prediction variances. The third row corresponds to the use of only hard data, and the fourth row corresponds to the use of both hard and soft data. The total number of observations used in each prediction procedure is also listed in the first two rows of each table.

Variable	**Mean**
Predicted Values-Hard (N=744)	18.39
Predicted Values-Hard and Soft (N=897)	18.24
Prediction Variance-Hard	31.39
Prediction Variance-Hard and Soft	30.84

Table 2.8: Summary of 2004 predicted nitrate concentrations (mg/L) from USDA-CSREES's CEAP study

Variable	**Mean**
Predicted Values-Hard (N=671)	18.18
Predicted Values-Hard and Soft (N=849)	18.19
Prediction Variance-Hard	31.04
Prediction Variance-Hard and Soft	30.56

Table 2.9: Summary of 2005 predicted nitrate concentrations (mg/L) from USDA-CSREES's CEAP study

Variable	Mean
Predicted Values-Hard (N=625)	17.87
Predicted Values-Hard and Soft (N=841)	17.61
Prediction Variance-Hard	31.60
Prediction Variance-Hard and Soft	30.55

Table 2.10: Summary of 2006 predicted nitrate concentrations (mg/L) from USDA-CSREES's CEAP study

2.9.3 Conclusions

In all three years, the average variance associated with the predicted values was smaller when both the hard and soft data were used in prediction. The additional information that the soft data provided resulted in this reduction in the standard error. In other words, incorporating this data into the prediction process led to more precise results.

The increase in precision became increasingly apparent when hard data points were limited. The smallest number of hard data points among the three years was recorded in 2006. Only 625 observations were measured, and an additional 216 were added as soft data points. Thus, approximately 35% of the data used in prediction were soft data. The average variance when only the hard data were used was 31.60 while the average variance when both the hard and soft data were used was 30.55, a difference of 1.05. This was the largest difference among the three years.

On the other hand, the largest number of hard data points among the three years was recorded in 2004 with 744 observations. An additional 153 points were available as

soft data so approximately 21% of the data were soft. The average variance when only the hard data were used was 31.39 while the average variance when both the hard and soft data were used was 30.84, a difference of only 0.55. The increase in precision was not as evident due to the larger number of hard data points.

2.9.4 Kriging Maps

Two ordinary kriging prediction maps were produced for each year using ArcGIS Version 9.2 (ESRI, 2006). The first map displays the results obtained from kriging with only the hard data while the second map displays the results from using the hard data along with the previous year's data as soft data in weighted kriging. Therefore, the more precise predictions are displayed in the second map. In 2004 and 2006, the mean nitrate concentrations based only on the hard data were higher, as indicated by darker colors on the maps. In 2005, the means were similar so the differences between the two maps are subtle.

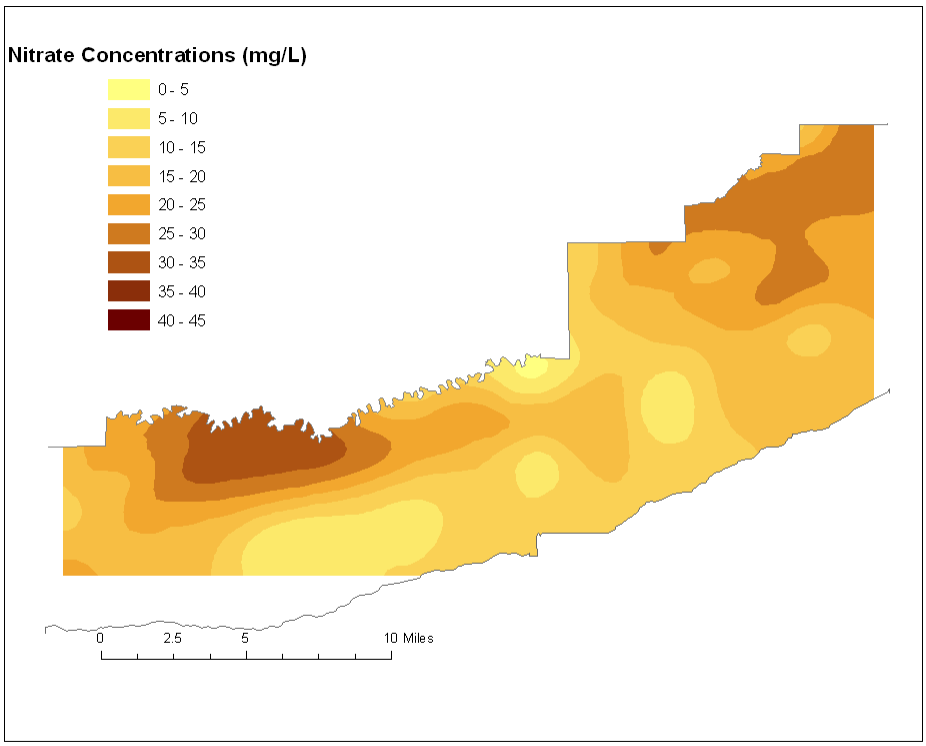

Figure 2.6: Kriging map for 2004-hard

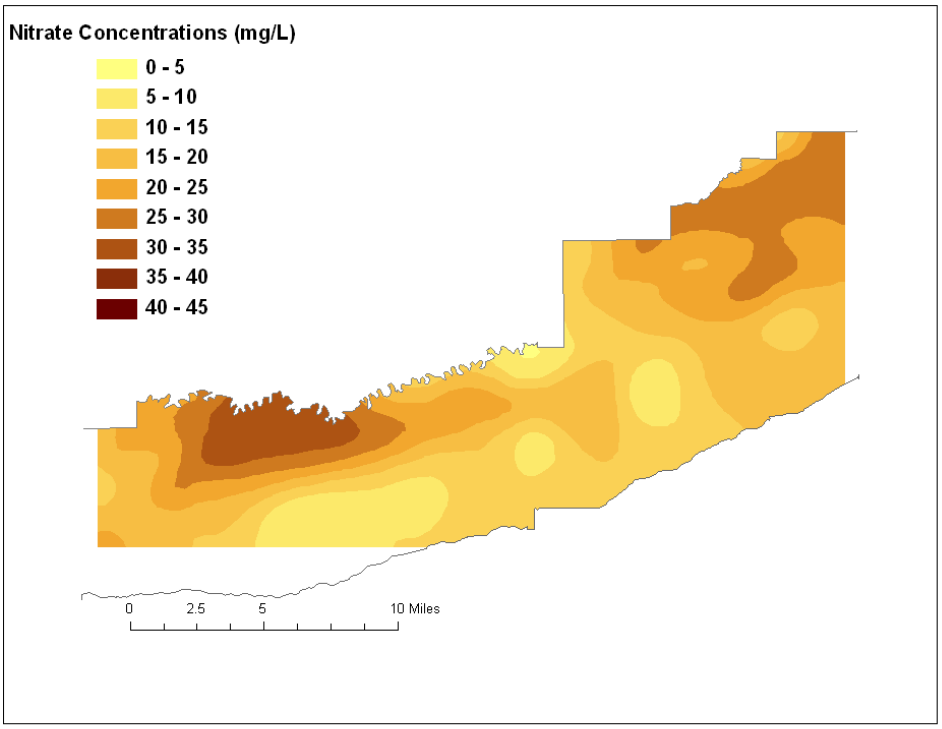

Figure 2.7: Kriging map for 2004-hard and soft

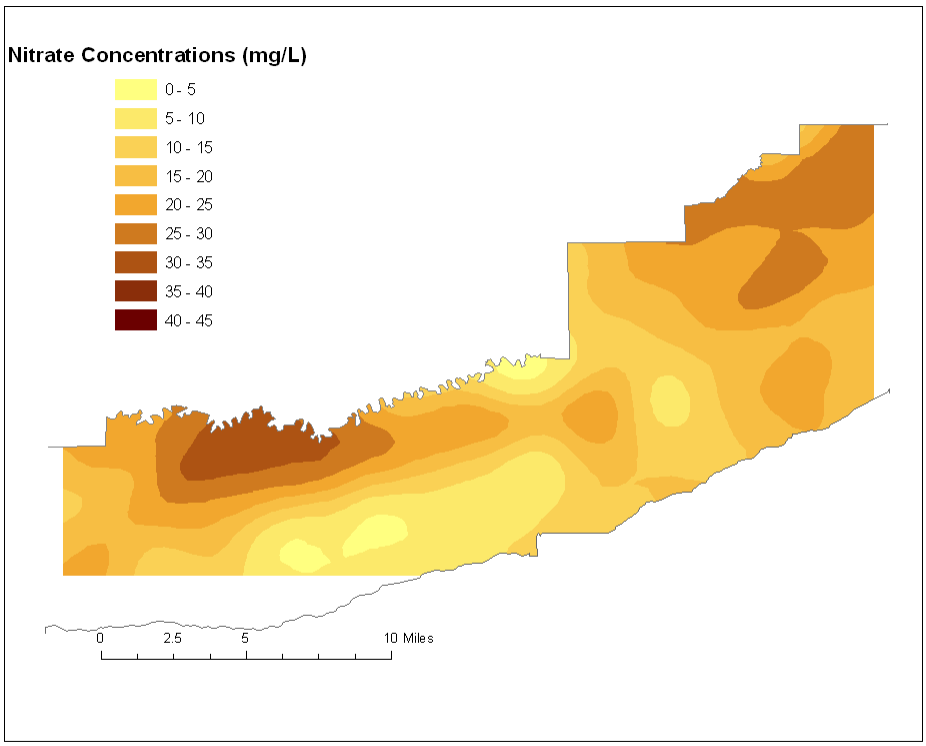

Figure 2.8: Kriging map for 2005-hard

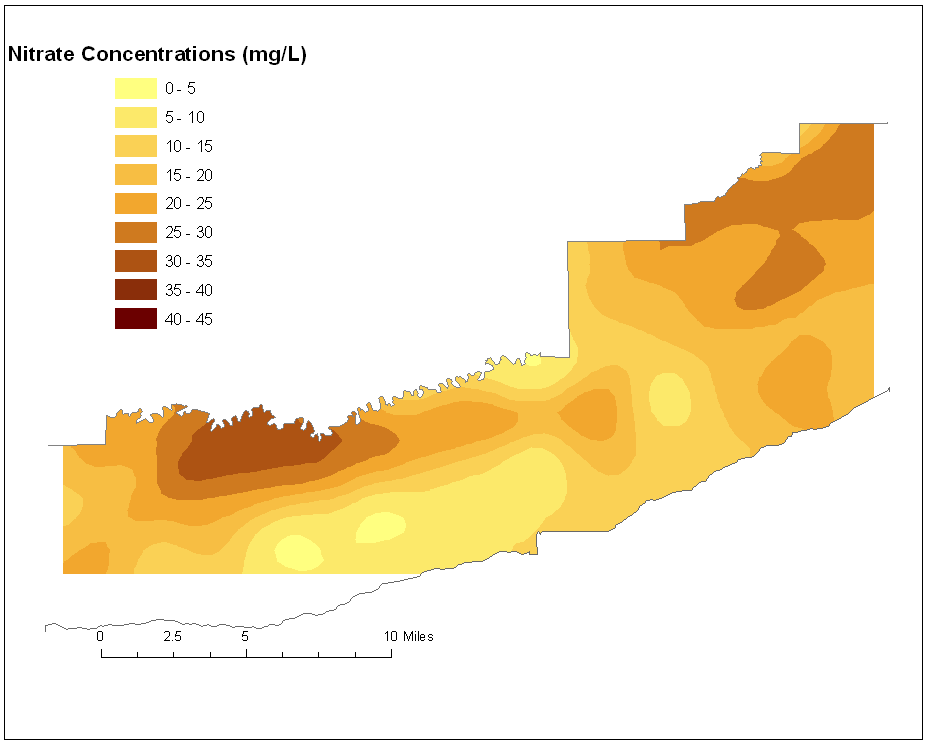

Figure 2.9: Kriging map for 2005-hard and soft

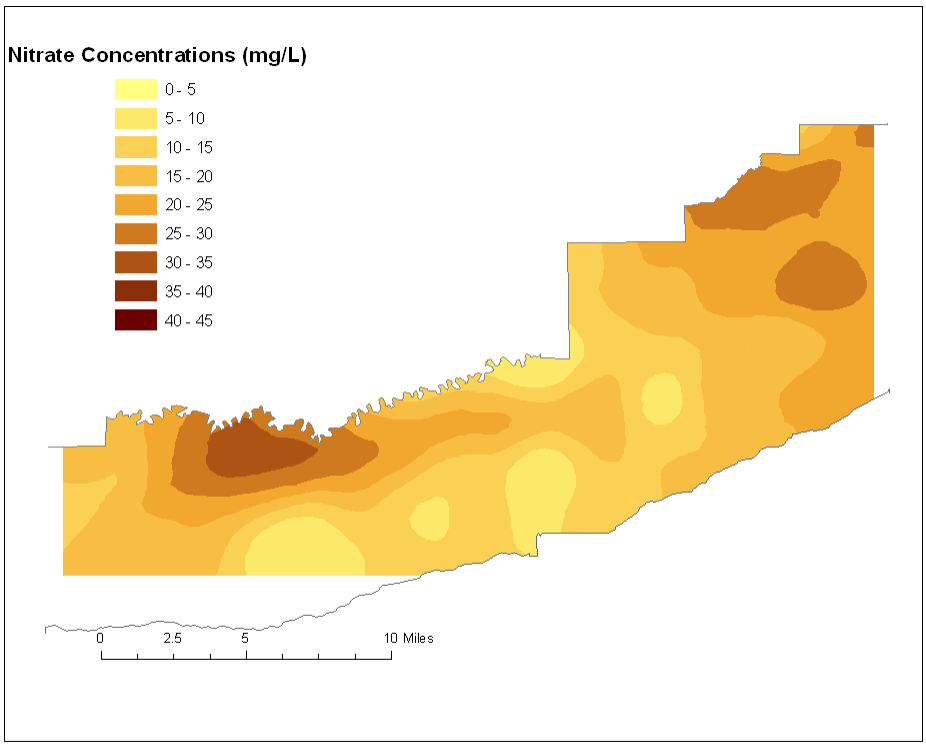

Figure 2.10: Kriging map for 2006-hard

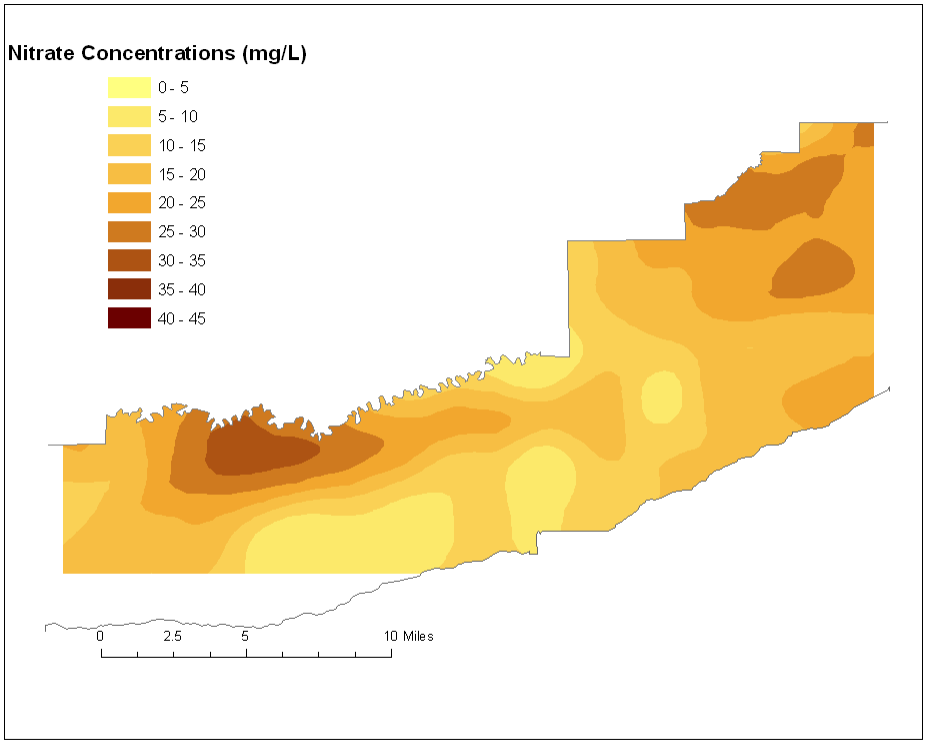

Figure 2.11: Kriging map for 2006-hard and soft

2.10 References

Christakos, G. (1990). A Bayesian/maximum-entropy view to the spatial estimation problem. *Mathematical Geology, 22* (7), 763-777.

Christakos, G. & Li, X. (1998). Bayesian maximum entropy analysis and mapping: A farewell to kriging estimators? *Mathematical Geology, 30* (4), 435-462.

Clark, I. & Harper, W. V. (2000). *Practical geostatistics*. Columbus, Ohio: Ecosse North America.

Cressie, N. (1991). *Statistics for spatial data*. New York: John Wiley & Sons.

ESRI. (2001). *Using ArcGIS geostatistical analyst*. Redlands, CA: ESRI.

ESRI. (2006). *ArcGIS desktop help 9.2*. Redlands, CA: ESRI.

Isaaks, E. H. & Srivastava, R. M. (1989). *Applied geostatistics.* New York: Oxford University Press.

Journel, A. G. (1986). Constrained interpolation and qualitative information-The soft kriging approach. *Mathematical Geology, 18* (3), 269-286.

Journel, A. G. & Huijbregts, Ch. J. (1978). *Mining geostatistics*. New York: Academic Press.

Lee, Y.-M., & Ellis, J. H. (1997). On the equivalence of kriging and maximum entropy estimators. *Mathematical Geology, 29* (1), 131-152.

Liedtke, M., Marx D., & Kachman S. (2009, January). Incorporating soft data in the kriging equations. Paper presented at the 8th Annual Hawaii International Conference on Statistics, Mathematics and Related Mathematics, Honolulu, Hawaii.

Matheron, G. (1963). Principles of geostatistics. *Economic Geology and the Bulletin of the Society of Economic Geologists*, 58, 1246-1266.

Olea, R. A. (2006). A six-step practical approach to semivariogram modeling. *Stochastic Environmental Research Risk Assessment, 20*, 307-318.

SAS Institute. (2008). *SAS online doc, Version 9.2*. Cary, NC: SAS Institute.

Schabenberger, O. & Gotway, C. A. (2005) *Statistical methods for spatial data analysis*. Boca Raton, Florida: Chapman & Hall/CRC.

Serre, M.L. (1999). Environmental spatiotemporal mapping and ground water flow modelling using the BME and ST methods (Ph.D. Dissertation, University of North Carolina at Chapel Hill, 1999).

Serre, M.L. (2007, July). *Introduction to Bayesian maximum entropy*. Paper presented at the BME workshop sponsored by the Department of Statistics, University of Nebraska-Lincoln.

Chapter 3 Weighted Kriging vs. Bayesian Maximum Entropy: Gaussian

3.1 Introduction

Although ordinary kriging provides the best linear unbiased predictor, it does have prediction limitations. In particular, kriging does not provide a framework to incorporate data of differing precisions. Chapter 2 focused on overcoming this limitation by incorporating soft data into the kriging equations by means of weighted kriging. A simulation study illustrated that weighted kriging yields more desirable fit statistics than traditional kriging techniques. However, the methodology which is commonly used to incorporate data of differing precision is called Bayesian Maximum Entropy (BME). This is a spatial/temporal mapping method capable of accounting for general knowledge and soft information (Kolovos, 2001). To compare these competing approaches, the data sets from the simulation in Chapter 2 were used in this chapter to compute BME predictions and their corresponding standard errors. Thus, probabilistic soft data in the form of the Gaussian distribution were used. BME validation statistics were then calculated and compared to the corresponding fit statistics obtained from weighted kriging.

3.2 Bayesian Maximum Entropy

The Bayesian Maximum Entropy (BME) approach was introduced in 1990 by George Christakos in a work entitled “A Bayesian/maximum-entropy view to the spatial

estimation problem." BME, unlike the long-existing prediction techniques, has the ability to combine data from various sources and of varying quality for spatiotemporal prediction (Christakos, 1990). In other words, BME has the power to incorporate soft data in a spatial analysis. More specifically, Christakos (1990) summarizes BME as an approach which accounts for prior knowledge, produces a posterior probability with minimum uncertainty, avoids Gaussian and unbiasedness assumptions, and yields results similar to those from well-established techniques when the same information is used. This methodology has been applied to a number of real-world environmental health studies (See Choi, Serre, Christakos, 2003; Christakos, 2009; Law et al., 2006; Savelieva, Demyanov, Kanevski, Serre, Christakos, 2005; Serre, Kolovos, Christakos, Modis, 2003).

According to Serre (1999), "the double epistemological goal of BME is informativeness (prior information maximization given general knowledge) and cogency (posterior probability maximization given specificatory knowledge)" (pg.3). To obtain this goal, BME progresses through three major stages of analysis. In the first stage, the prior stage, the basic assumptions are given and the form of a prior probability density function is derived such that its entropy is maximized subject to the general knowledge available (Serre, 1999).

The second stage, called the meta-prior or pre-posterior stage, considers the specificatory knowledge composed of both the hard and soft data (Serre, 1999). The third and final stage is the integration or posterior stage (Serre, 1999). Both knowledge bases are considered in this stage, and the goal is to maximize the posterior probability given both the general knowledge and the specificatory knowledge (Serre & Christakos,

1999). Using Bayesian conditionalization to update the prior probability distribution function with respect to the specific data collected, the posterior probability density function is derived (Orton & Lark, 2007a). This posterior distribution provides the BME prediction (Orton & Lark, 2007a).

In certain situations, kriging and BME produce identical results. When only hard data are used and the local mean is known, BME predictions are the same as simple kriging predictions (Christakos & Li, 1998; Orton & Lark, 2007b). Lee and Ellis (1997) also showed that if the random field is assumed to be second-order stationary or Gaussian, then the simple kriging and maximum entropy predictions are equivalent. In addition, when only hard data are used and the mean is assumed to be given by an unknown constant, BME predictions are the same as those from ordinary kriging (Orton & Lark, 2007b).

3.3 The SEKS-GUI software library

The Spatiotemporal Epistemic Knowledge Synthesis-Graphical User Interface or SEKS-GUI package combines the Bayesian Maximum Entropy library (BMElib) and the Generalized BME library (Kolovos, Yu, & Christakos, 2006). Since the primary focus of this chapter is to compare BME to weighted kriging, the BMElib was used for space modeling, estimation, and mapping. This library processes detrended, normally distributed data sets and allows for a detailed exploratory data analysis (Yu, Kolovos, Christakos, Chen, Warmerdam, & Dev, 2007). In addition, the user can model correlations by fitting covariance models to the data (Yu et al., 2007).

Figure 3.1 summarizes the modeling and mapping phases of the SEKS-GUI (Kolovos et al., 2006).

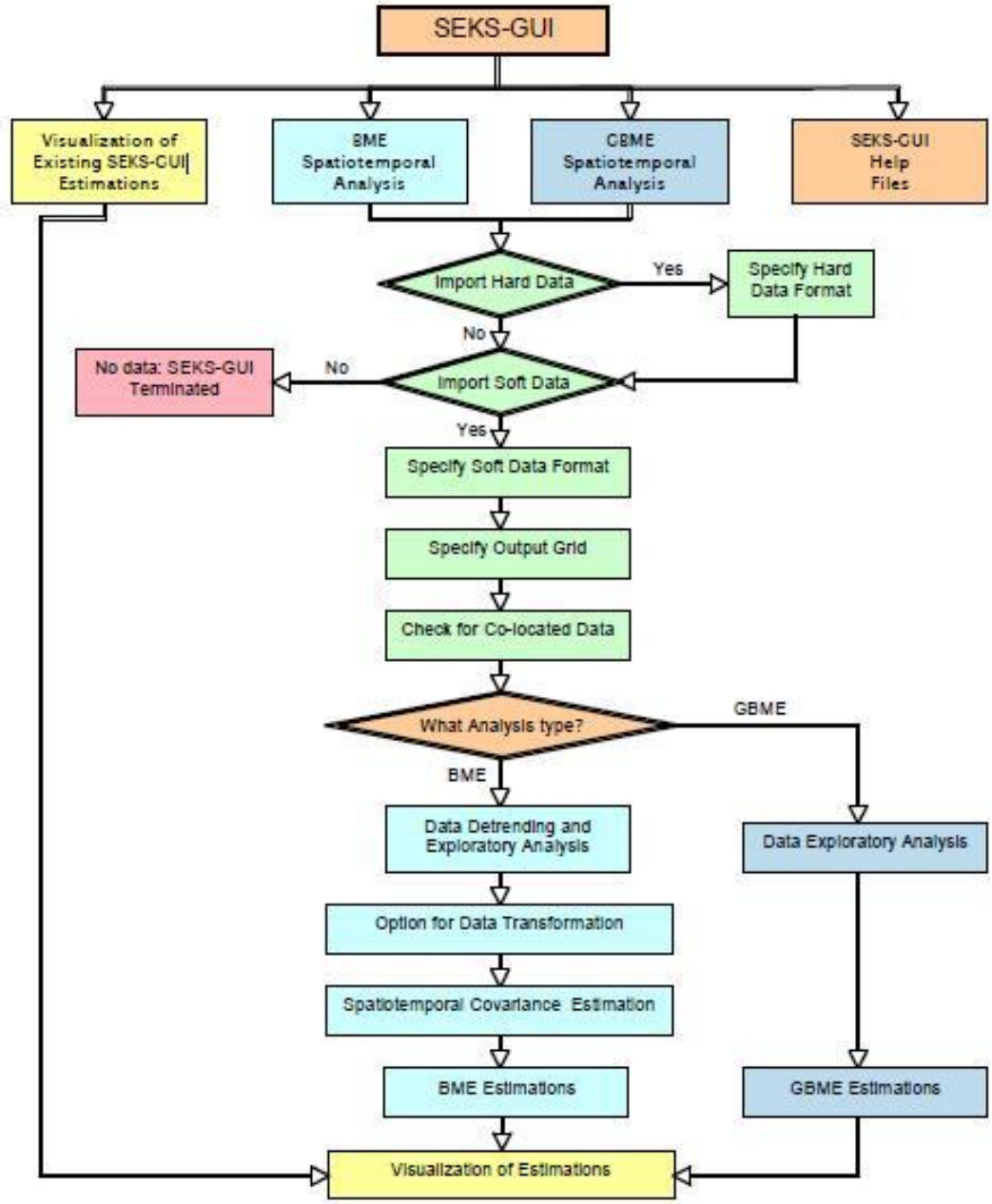

Figure 3.1: Flowchart of SEKS-GUI

3.4 Simulation Study

The purpose of this study was to compute BME predictions and their corresponding prediction standard errors using the SEKS-GUI. This was done by implementing the SEKS-GUI package in Matlab Version 7.3.0 (2006). The BME predicted values and standard errors were used to calculate the validation statistics provided in Section 2.6. The means of these statistics were then compared to the corresponding statistics from weighted kriging.

In this section, one of the data sets with a simulated range of 15 and 10% soft data is used to illustrate the sequence of interactive screens provided by the SEKS-GUI procedure. The first step is shown in Figure 3.2 and corresponds to choosing the appropriate task within the SEKS-GUI package. For this study, the BME Spatiotemporal Analysis procedure was selected.

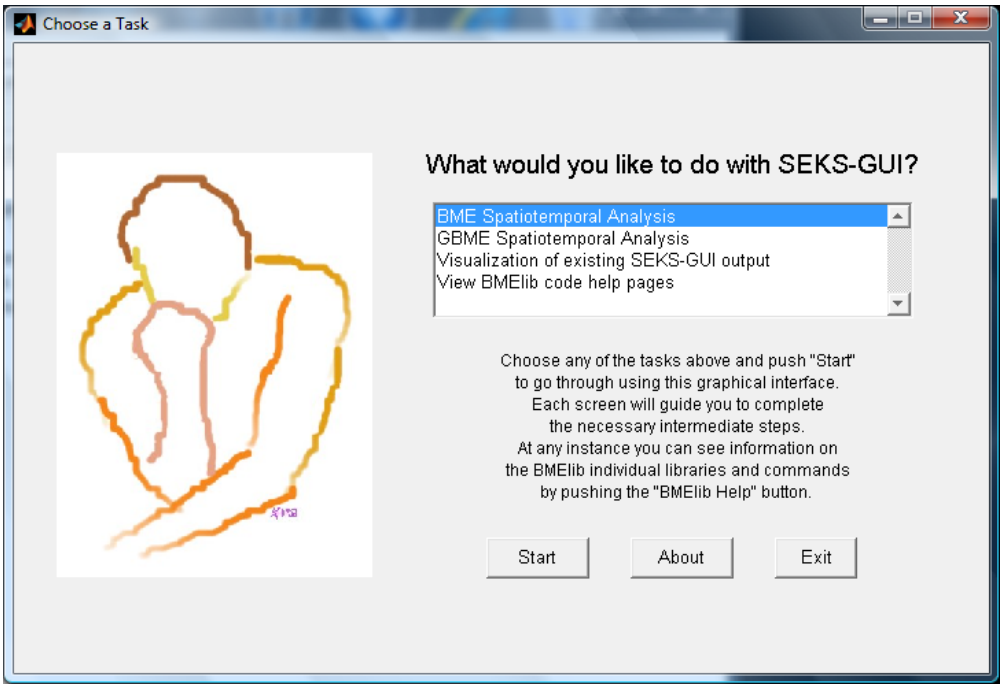

Figure 3.2: A screenshot of task options in SEKS-GUI

The next step was to enter the hard and soft data into the system. Recall, a spherical spatial floor was simulated on a 40 by 40 grid (1600 points) with a sill of 1.0 and a nugget of 0, and a range of either 15 or 30. From the 1600 points, 400 were randomly selected to be observed values and the remaining 1200 were used for validation. As outlined in Table 2.1, either 10% or 50% of the observed values were randomly chosen to be soft data. These points became soft by adding an independent Normal component with a mean of 0 and variance of 0.5. In order to make a fair comparison, the hard and soft data files used to obtain the BME predictions were the same as those used to obtain the weighted kriging predictions in Chapter 2.

Figure 3.3 shows the screen which allows the user to select the appropriate hard data file. At this time, the user must also specify if the study is purely spatial or if it is both spatial and temporal. For this study, it was appropriate to check the box indicating a space-only domain.

Figure 3.3: A screenshot of hard data selection in SEKS-GUI

After selecting the appropriate hard data file, the user must specify which columns in the file contain the spatial coordinates and which column contains the hard data values. This step is shown in Figure 3.4.

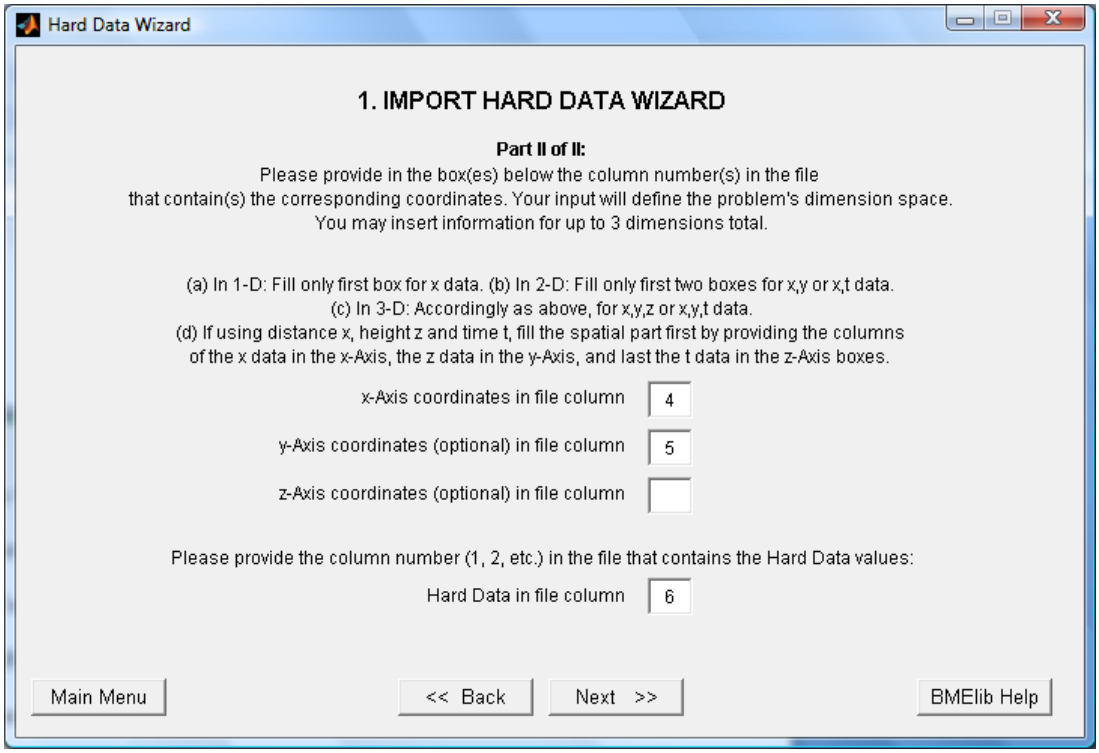

Figure 3.4: A screenshot of column selection in SEKS-GUI

As mentioned in Chapter 2, there are two types of soft data. Interval soft data are provided in terms of a lower and upper bound, whereas probabilistic soft data are provided in the form of a probability density function (pdf) (Serre, 2007). The SEKS-GUI accepts probabilistic soft data with fully described pdf characteristics, including data in the form of Gaussian, uniform, or triangular distributions, and those with user described pdfs (Kolovos et al., 2006).

In this study, probabilistic soft data in the form of the Gaussian distribution were used. Thus, each soft data point consisted of its spatial coordinates (xA, yA) and its mean and variance. The mean for each data point corresponded to the variable which resulted after the addition of the N(0, 0.5) component, and a variance equal to 1.5 was

specified for each soft data point. The reasoning for this variance value was because the variance of the hard data was equal to 1, i.e., the sill, and thus, the variance of the soft data was equal to the sill plus the added variance, i.e., 1+0.5 = 1.5. The two screens associated with the importation of the soft data are shown in Figures 3.5 and 3.6.

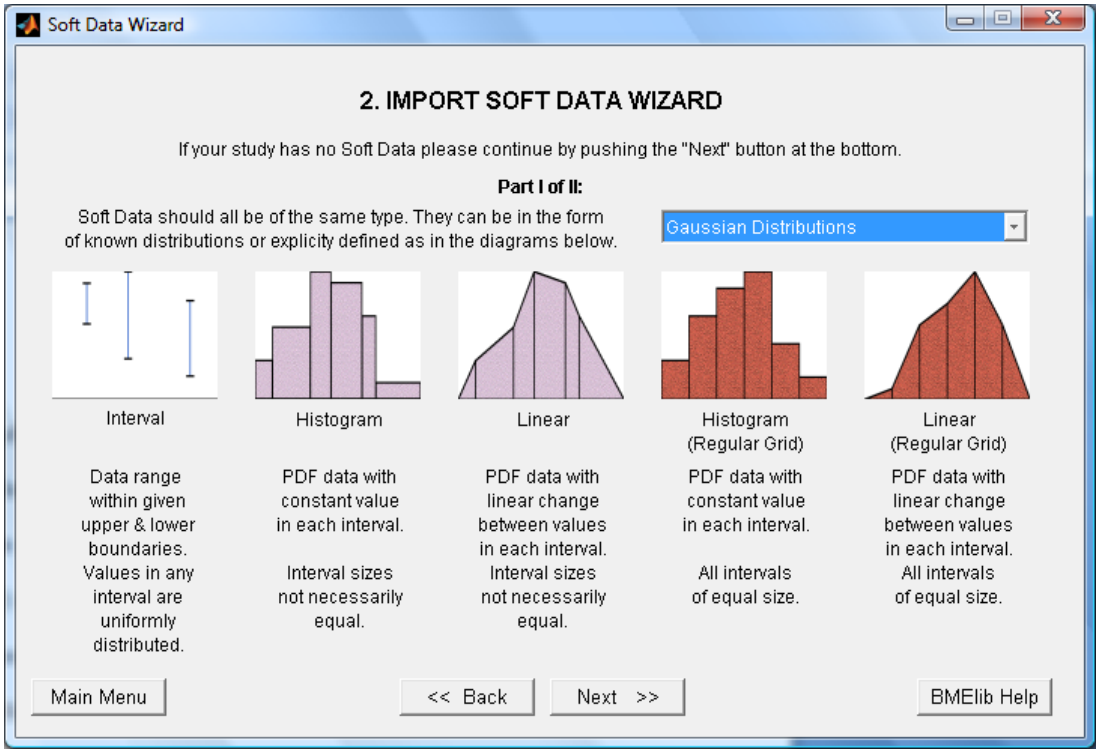

Figure 3.5: A screenshot of soft data types in SEKS-GUI with Gaussian selected

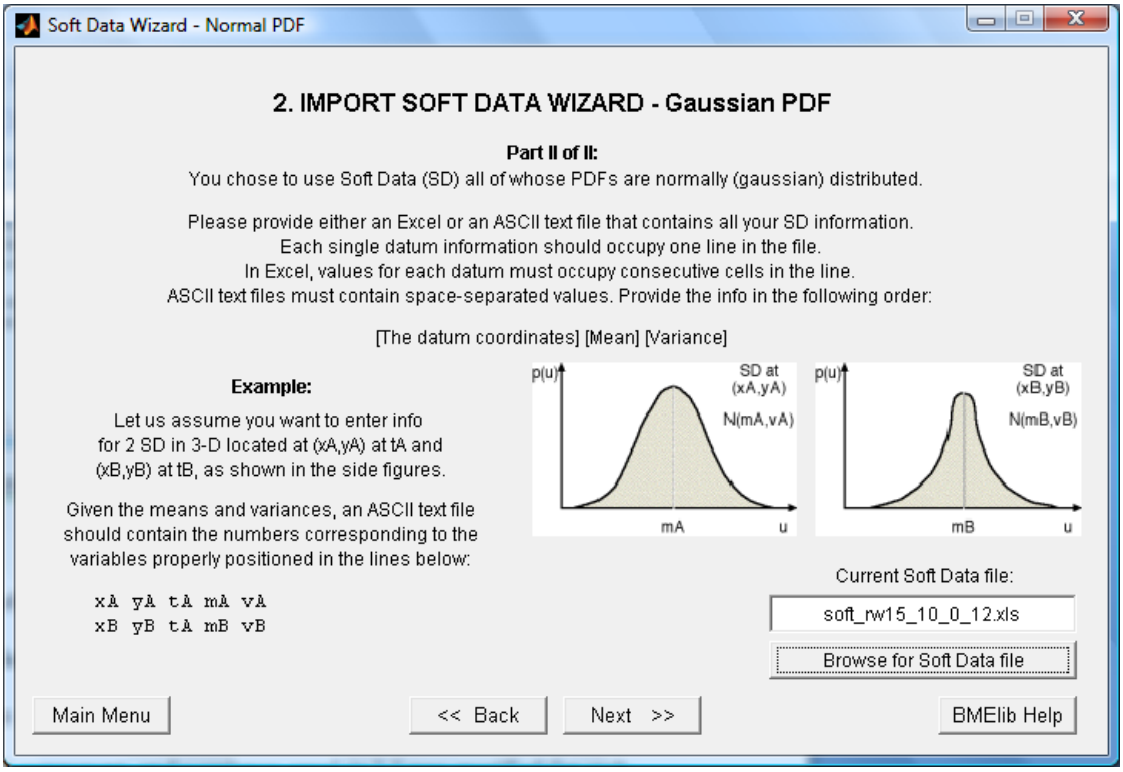

Figure 3.6: A screenshot of importing soft data with Gaussian distribution in SEKS-GUI

The next step requires the user to define the locations where predicted values are to be obtained. This grid file must be formatted according to one of three options. The second option, option B, was used in this study. This option specified that the grid file contained grid limits and the number of nodes in each dimension. The grid file for this study was an Excel file with two rows and three columns. The first row corresponded to the first spatial coordinate, X, and the second row corresponded to the second spatial coordinate, Y. The first column represented the lower limit for each spatial coordinate, column two represented the number of nodes (points) to be predicted, and column three represented the upper limit for the spatial coordinate. Thus, the grid file was formatted like Table 3.1. As shown in Figure 3.7, this step also allows the user to select whether or

not they are only mapping positive values. For this study, only positive values were appropriate.

1	40	40
1	40	40

Table 3.1: Output grid file used in SEKS-GUI

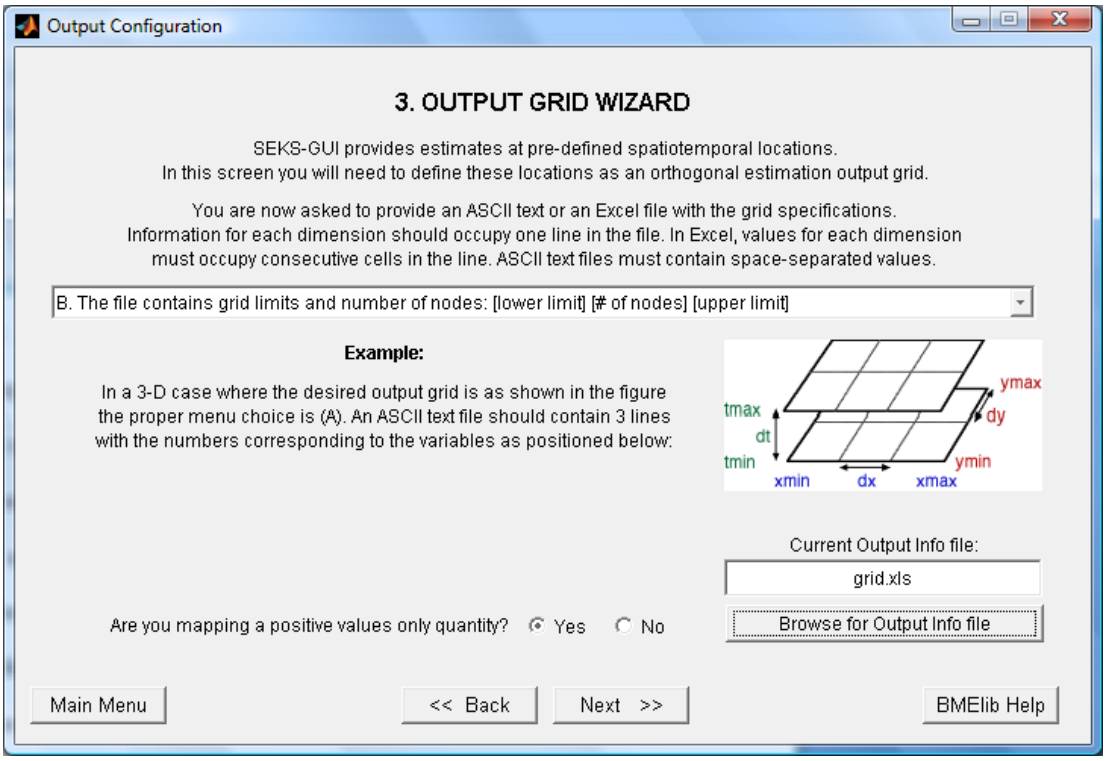

Figure 3.7: A screenshot of output grid selection in SEKS-GUI

After the data sets and output grid are entered into the system, the SEKS-GUI procedure enters the "Exploratory Analysis" phase. This phase consists of three steps. The first step, shown in Figure 3.8, checks the hard and soft data files for duplicates, i.e., multiple observations at the same location. If duplicates are present, it can affect the

covariance analysis. However, this was not a concern in this study because observations of this type were not created in the simulation process.

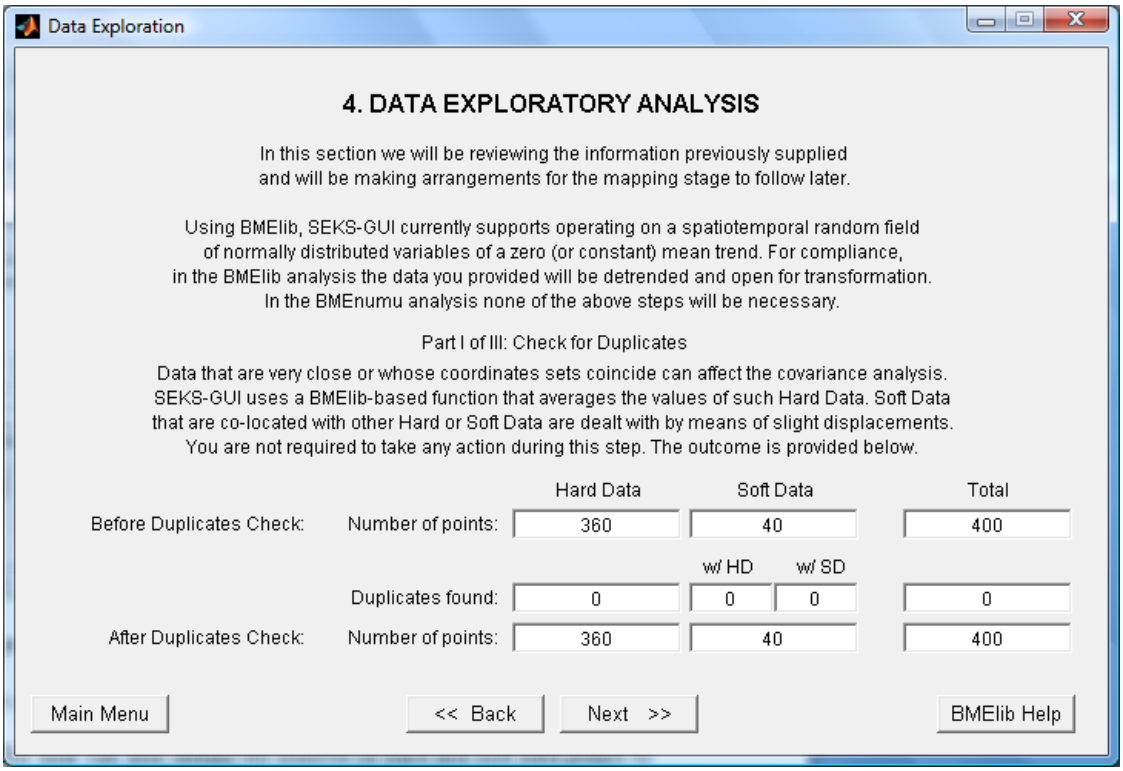

Figure 3.8: A screenshot of the data check in SEKS-GUI

The next step in the SEKS-GUI procedure is to remove any trends in the data and check the detrended data for normality. If appropriate, a transformation of the data can be performed. The previous two actions were not necessary in this study because each simulated data set followed a nearly normal distribution (See Figures 3.9-3.10).

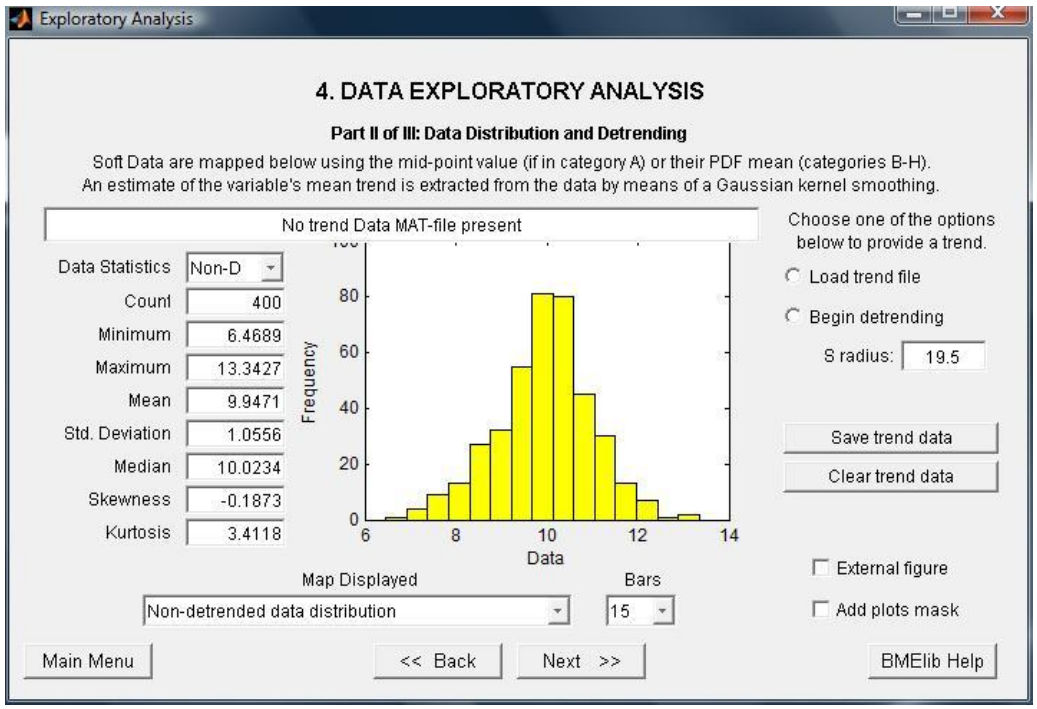

Figure 3.9: A screenshot of the detrending screen in SEKS-GUI

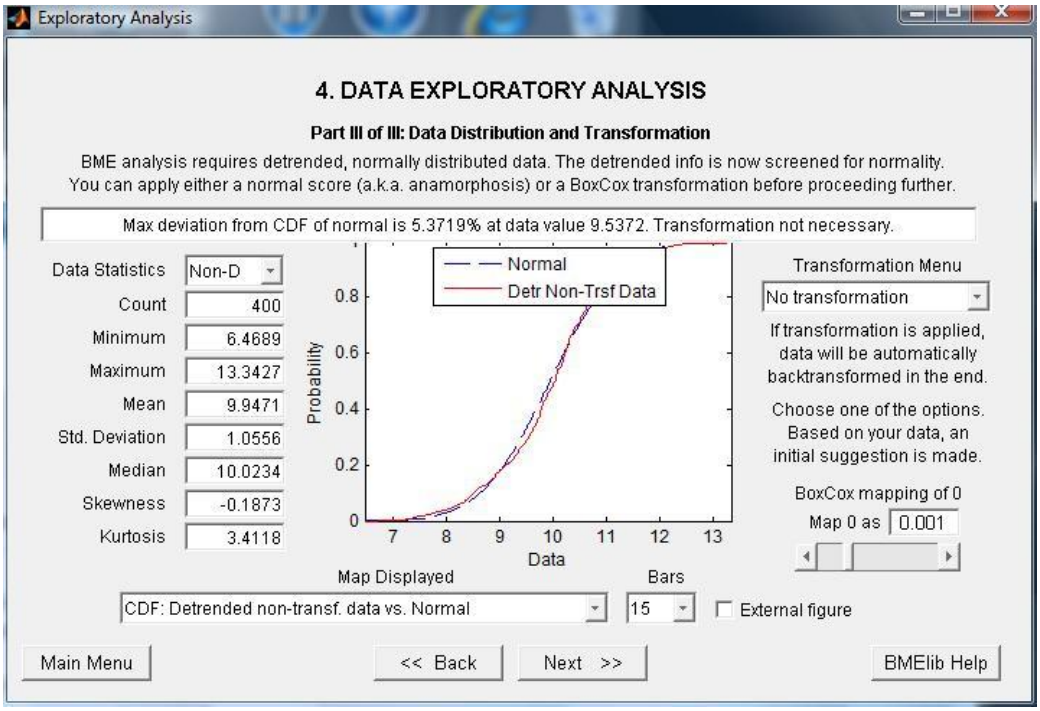

Figure 3.10: A screenshot of the data transformation screen in SEKS-GUI

The next phase of the procedure is the "Covariance Analysis" (See Figure 3.11). In this phase, the spatial correlation patterns in the data are modeled through a particular covariance function (Kolovos et al., 2006). This can be done by splitting the prediction field into sub-grids or alternatively, by treating the prediction field as one solid neighborhood (Kolovos et al., 2006). The latter approach, one grid with one covariance function, was chosen for this study.

To initiate the calculations on the BMElib experimental covariance, the user can click on the "Get experimental" button after the correlation range and lag parameters are set. Then a covariance model is fit to the experimental covariance information. In the SEKS-GUI procedure, the model fit is based purely on visual inspection and can be adjusted by changing the sill and range parameters (Kolovos et al., 2006).

In this study, a Spherical model was always chosen and the sill parameter was left at its default value (approximately 1). The range, however, was adjusted in order to provide the best fit. For the data sets with a simulated range of 15, the range was set to 15 in the SEKS-GUI package. However, when the simulated range was 30, a range of 30 in the SEKS-GUI package did not provide the best visual fit. In all instances, the best fit corresponded to a range between 15 and 25.

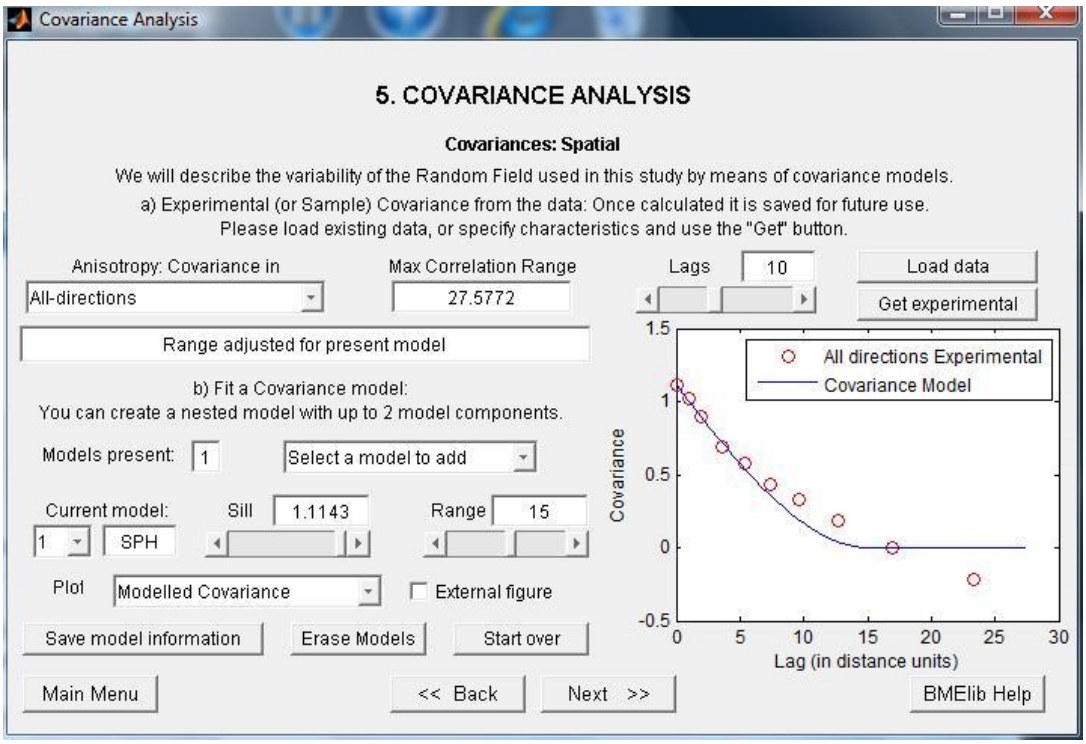

Figure 3.11: A screenshot of the covariance analysis stage in SEKS-GUI

Next, the user selects the type of estimation to be performed by BMElib. The option chosen in this study was "BME Moments (calculates the BME Mean, error var., and skewness)." The user can also define the number of hard and soft data points to be used for prediction. In the case where 10% of the data were soft, a maximum of 18 hard data points and 2 soft data points were used in prediction. These numbers were chosen because the SAS® weighted kriging program to which the SEKS-GUI procedure is being compared used the 20 nearest neighbors, and thus, these numbers made the procedures as similar as possible. When 50% of the data were soft, a maximum of 5 hard and 5 soft data points were used in prediction. Initially, in an attempt to stay consistent with the SAS® program, a maximum of 10 hard and 10 soft data points were used in prediction when 50% of the data were soft. However, a warning message appeared that the time

required for computations increases, in general, exponentially with the amount of soft data used, and therefore, recommended a maximum of 3-4 soft data points (Kolovos et al., 2006). If the message was ignored, the prediction of the 1600 points took over 8 minutes as opposed to approximately 2 minutes when only 2 soft data points were used. In addition, when the prediction was complete, the following message appeared: "Unacceptable results in estimations!" This implied that a predicted value was not calculated at some of the x, y coordinates. Due to these problems, a maximum of 5 hard and 5 soft data points were used in prediction to keep the procedures as similar as possible.

Figures 3.12-3.14 are screen captures of the SEKS-GUI package as it progresses through estimation and the final visualization phase. Figure 3.13 is a map of the mean of the estimation posterior probability density function (pdf) at each output grid node, and Figure 3.14 is a map of the standard deviation of the estimation posterior pdf at each output grid node. These maps provided the data to calculate the validation statistics.

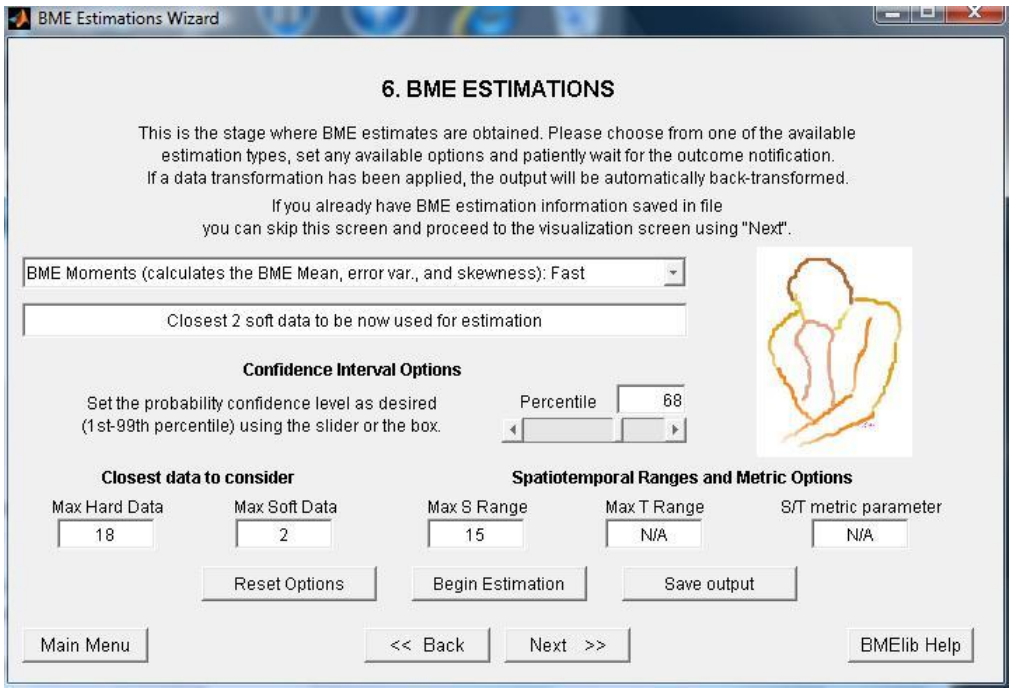

Figure 3.12: A screenshot of the prediction phase in SEKS-GUI

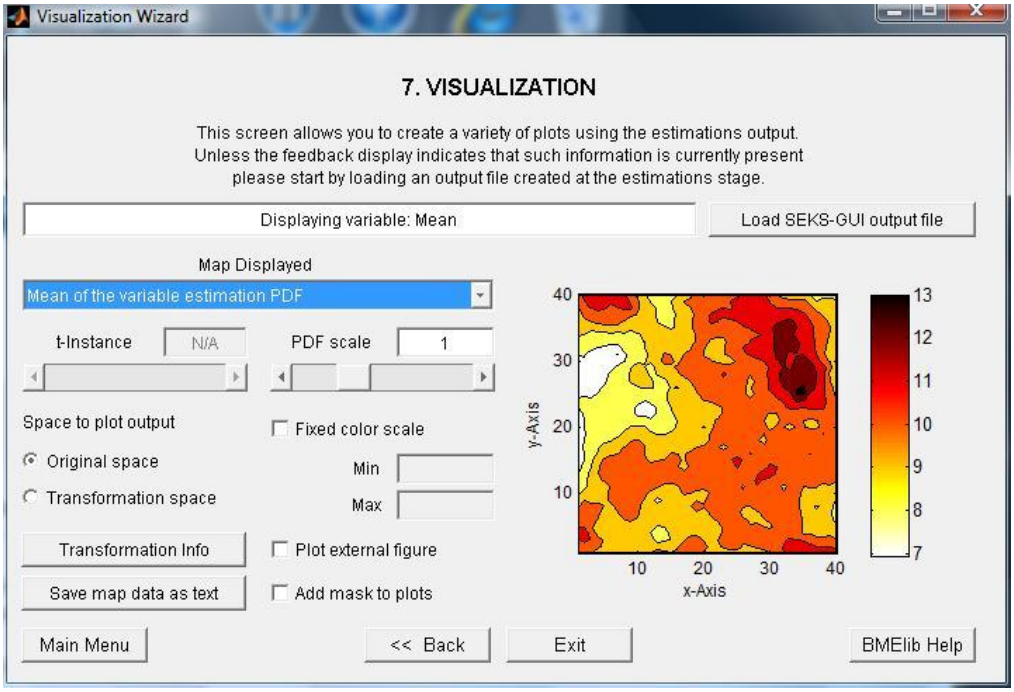

Figure 3.13: A screenshot of the predicted means in the visualization phase in SEKS-GUI

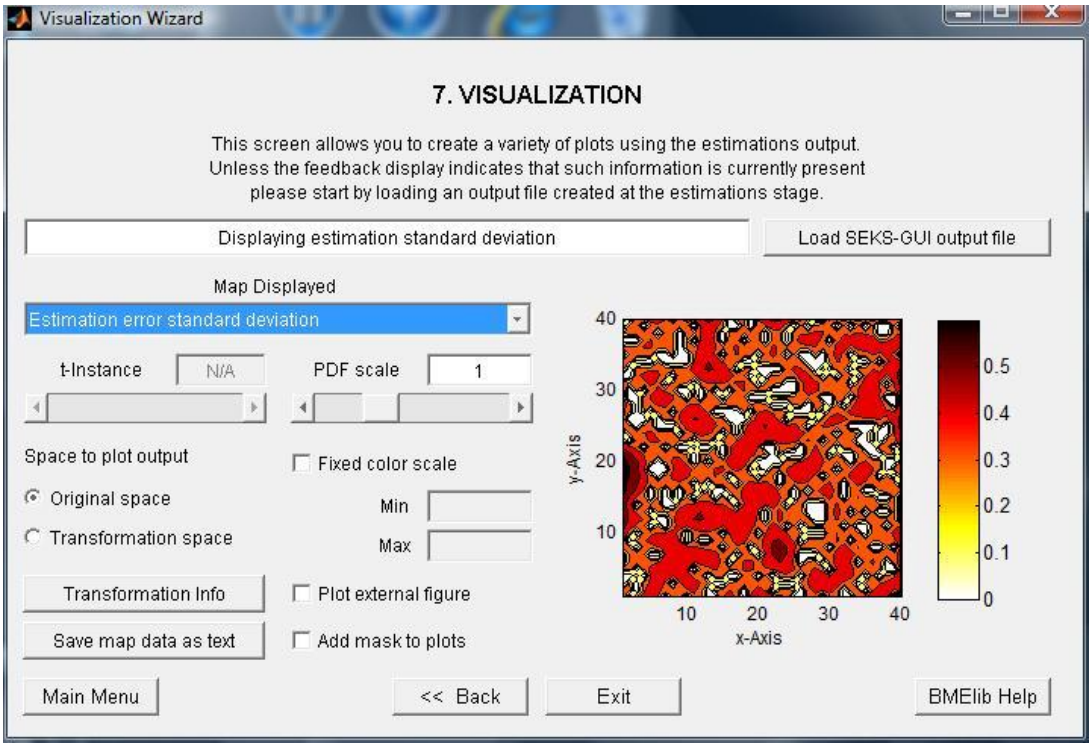

Figure 3.14: A screenshot of the prediction standard errors in the visualization stage in SEKS-GUI

.

3.5 Results

Tables 3.3-3.6 summarize the means of the fit statistics obtained from the SEKS-GUI BME Spatiotemporal analysis and from the weighted kriging procedure. These two procedures are denoted by **BME** and **Weighted**, respectively, in the tables. The fit statistics defined in Section 2.6 were used to compare the two prediction techniques.

	RMSE	AVAR	SME	ABSMPE	RMSSE
BME	0.4822	0.1417	-0.2352	0.3696	1.2851
Weighted	0.3738	0.2545	-0.0007	0.2954	0.7618

Table 3.2: Fit statistics obtained from BME and weighted kriging with a range=15 and 10% Gaussian soft data

	RMSE	AVAR	SME	ABSMPE	RMSSE
BME	0.6805	0.2415	-0.3435	0.5631	1.4842
Weighted	0.4563	0.3000	-0.0051	0.3267	0.8029

Table 3.3: Fit statistics obtained from BME and weighted kriging with a range=15 and 50% Gaussian soft data

	RMSE	AVAR	SME	ABSMPE	RMSSE
BME	0.3845	0.0999	-0.2329	0.2899	1.2248
Weighted	0.2975	0.1336	-0.0024	0.2146	0.9018

Table 3.4: Fit statistics obtained from BME and weighted kriging with a range=30 and 10% Gaussian soft data

	RMSE	AVAR	SME	ABSMPE	RMSSE
BME	0.6470	0.2116	-0.3188	0.5163	1.4683
Weighted	0.2835	0.1570	0.0005	0.2239	0.7350

Table 3.5: Fit statistics obtained from BME and weighted kriging with a range=30 and 50% Gaussian soft data

3.6 Summary

Friedman's Chi-Square Test was used to determine if there was a significant difference between the two prediction procedures. For this nonparametric test, each simulated data set served as a block, and the prediction technique was the treatment. In all four cases with varying ranges and percentages of soft data and for each error statistic, Friedman's test resulted in a highly significant p-value based on an alpha level of 0.05. Thus, there was a significant difference between the procedures.

Recall that the RMSE and the AVAR should be small for a model which fits the data well. The SME and ABSMPE should be close to zero while the RMSSE should be close to one. Thus, with the exception of the AVAR, the means of the fit statistics from weighted kriging were always more desirable than those from BME. Although the AVAR produced by weighted kriging was larger in three of the four cases, the accuracy of this statistic is more important as it is possible that BME produced fictitiously lower prediction errors.

In addition to comparing the two prediction procedures, the four simulation cases with varying ranges and percentages of soft data were also compared. The difference between the two procedures appeared to be more noticeable when 50% of the data were soft rather than 10%. Furthermore, with the exception of the SME, the means resulting from the simulated data sets with a range of 30 were always smaller than those resulting from the data sets with a range of 15. Overall, the weighted kriging procedure performed better than BME.

3.7 Model Fitting

As mentioned in Section 3.3, in the "Covariance Analysis" phase of the SEKS-GUI procedure, a covariance model is fit to the experimental covariance information. The model fit is based purely on visual inspection by the user and can be adjusted by changing the sill and range parameters. In this study, a Spherical model was always chosen and the sill parameter was left at its default value (approximately 1). The range, however, varied depending on the range specified when the data were simulated. To show how important it is to properly fit the model, the simulated data sets with a range of 15 and 10% soft data were used in the SEKS-GUI package without setting the range to 15. Instead, the range was left at its default value, which varied between 3 and 5. Table 3.7 summarizes the results where **BME-Default** corresponds to leaving the range at its default value, and **BME-15** corresponds to setting the range equal to 15. Therefore, the **BME-15** results in Table 3.7 correspond to the **BME** results in Table 3.2.

Type	RMSE	AVAR	SME	ABSMPE	RMSSE
BME-Default	1.6778	0.3820	1.9995	1.2984	2.4922
BME-15	0.4822	0.1417	-0.2352	0.3696	1.2851

Table 3.6: Fit statistics obtained from BME with default range and BME with specified range=15 using data with simulation range=15, 10% soft data

Friedman's Chi-Square Test was used to determine if there was a significant difference between the two procedures. For each statistic, this nonparametric test resulted in a highly significant p-value based on an alpha level of 0.05. Thus, there was a

significant difference between the prediction techniques. Based on the results in Table 3.6, it is obvious that a poor fitting model caused a dramatic increase in the means of the fit statistics.

The SEKS-GUI BME Spatiotemporal analysis also allowed for prediction outside the range of the observed data. When the range of the spherical model was left at its default value, the predicted values were as low as 2, but the observed values ranged from approximately 8 to 12. Figures 3.15 and 3.16 display the results from two specific data sets (IQ 14 and 15) simulated with a range of 15 and with 10% soft data. On the y-axis are the predicted values obtained from the weighted kriging procedure. On the x-axis are the values obtained from the map of the "mean of the variable estimation PDF" from SEKS-GUI BME Spatiotemporal analysis when the range was left at its default value. The straight line that is formed by the points at x = y corresponds to the 400 observed values with zero prediction error. The graphs show that BME consistently produced smaller predictions than weighted kriging for the remaining 1200 points. In addition, the predicted values from the two procedures were weakly correlated with adjusted R^2 values of 0.30 and 0.19.

Predicted Values from Weighted Kriging vs. BME with Default Range Using Data 14

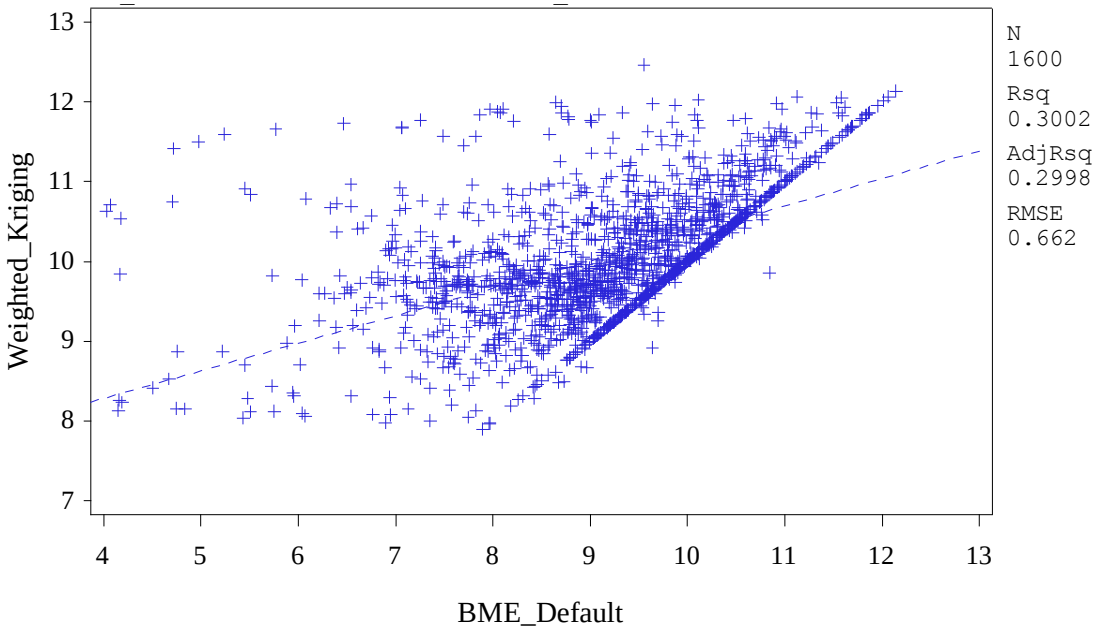

Figure 3.15: Plot of simulated data set IQ 14 of predicted values from weighted kriging against BME with default range

Predicted Values from Weighted Kriging vs. BME with Default Range Using Data 15

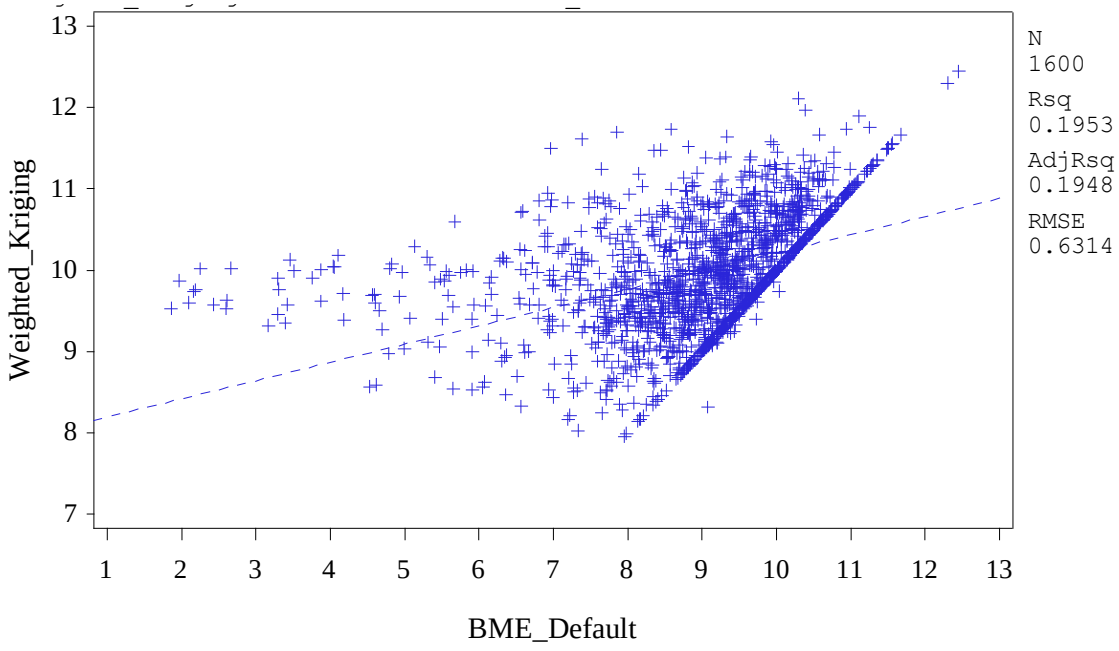

Figure 3.16: Plot of simulated data set IQ 15 of predicted values from weighted kriging against BME with default range

For comparison purposes, the same two data sets (range=15, 10% soft data, IQ 14 and 15) were used to produce Figures 3.17 and 3.18. These plots differ from those above because the x-axis values were obtained by setting the range equal to 15 in the BME procedure. In other words, the range which provided the best fit was selected. As in the previous plots, the y-axis values are the predicted values obtained using weighted kriging. Likewise, the straight line that appears at x = y corresponds to the 400 observed values with zero prediction error. It is clear that selecting an appropriate range resulted in highly correlated predicted values as the adjusted R^2 values were 0.89 and 0.87.

Predicted Values from Weighted Kriging vs. BME with Range=15 Using Data 14

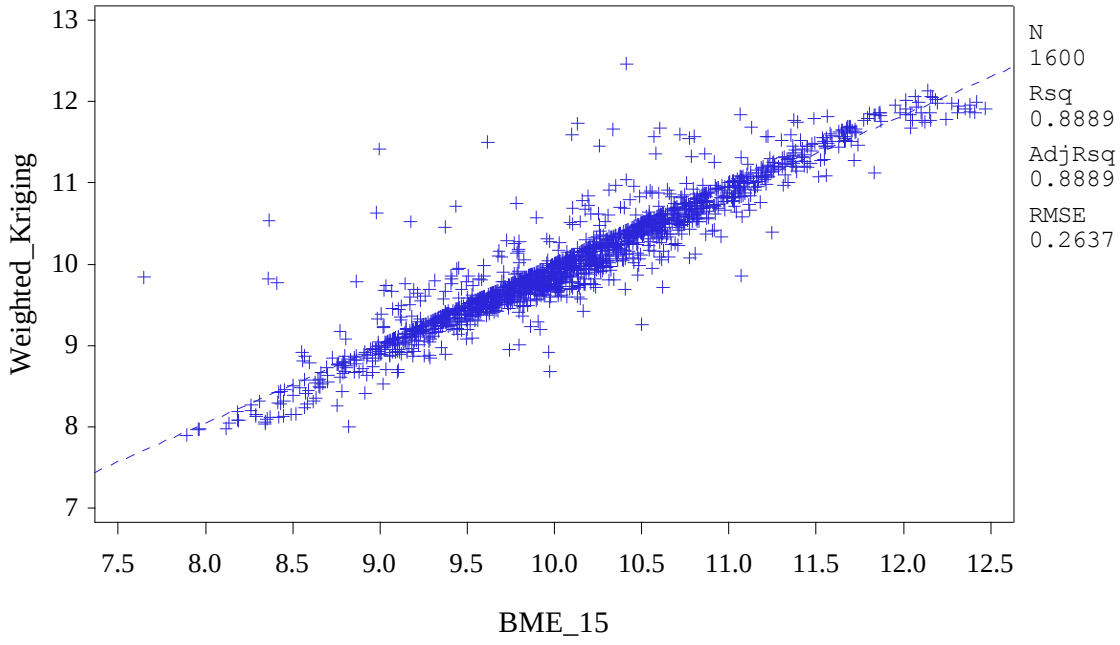

Figure 3.17: Plot of simulated data set IQ 14 of predicted values from weighted kriging against BME with specified range=15

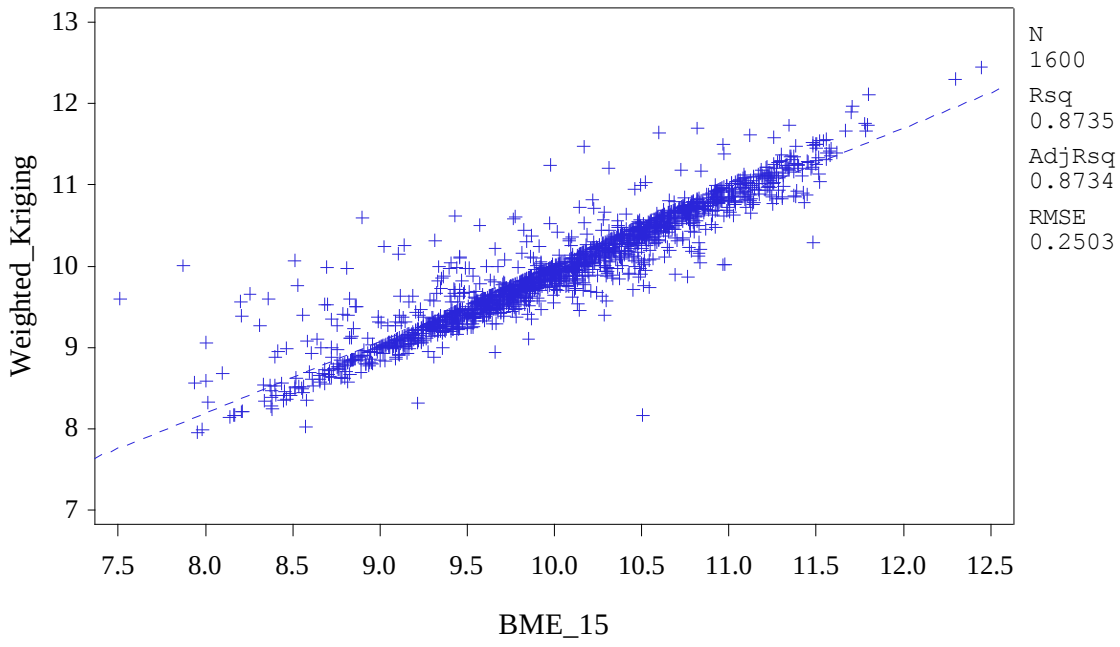

Figure 3.18: Plot of simulated data set IQ 15 of predicted values from weighted kriging against BME with specified range=15

3.8 Conclusions

Bayesian Maximum Entropy (BME) is a generalization of the well-established prediction techniques used in geostatistics (Christakos, 1990). This methodology has the ability to incorporate soft data into a spatial analysis in a systematic manner. In order to implement the methodology, the SEKS-GUI interactive software library can be used for space-time modeling, prediction, and mapping (Kolovos et al., 2006). In this chapter, the software library was used in a simulation study to compare weighted kriging to BME. The hard and soft data sets simulated in Chapter 2 were used in the SEKS-GUI package to obtain BME predictions and to produce a map of the mean of the estimation posterior probability density function (pdf) and a map of the standard deviation of the estimation

posterior pdf. Based on the data from these maps, the BME validation statistics were computed and determined to be less desirable than those obtained from weighted kriging. Furthermore, it was shown that the results obtained from the SEKS-GUI software library are extremely sensitive to the model parameters, and thus, it is crucial to fit the model accurately. Thus, although BME is a powerful method for spatial prediction, weighted kriging is a more robust procedure.

3.8 References

Choi, K. M., Serre, M. L., & Christakos, G. (2003). Efficient mapping of California mortality fields at different spatial scales. *Journal of Exposure Analysis and Environmental Epidemiology, 13* (2), 120-133.

Christakos, G. (1990). A Bayesian/maximum-entropy view to the spatial estimation problem. *Mathematical Geology, 22* (7), 763-777.

Christakos, G. (2009). *Epistematics: An evolutionary framework of real world problem-solving*. New York: Springer.

Christakos, G. & Li, X. (1998). Bayesian maximum entropy analysis and mapping: A farewell to kriging estimators? *Mathematical Geology, 30* (4), 435-462.

Kolovos, A. (2001). Computational investigations of the BME mapping approach and incorporation of physical knowledge bases (Ph.D. Dissertation. University of North Carolina at Chapel Hill, 2001).

Kolovos, A., Yu H.-L., & Christakos, G. (2006). *SEKS-GUI v.0.6 user manual.* San Diego: Department of Geography, San Diego State University.

Law, D. C. G., Bernstein, K. T., Serre, M. L., Schumacher, C. M., Leone P. A., Zenilman, J. M., et al. (2006). Modeling a syphilis outbreak through space and time using the Bayesian maximum entropy approach. *Annals of Epidemiology, 16* (11), 797-804.

Lee, Y.-M., & Ellis, J. H. (1997). On the equivalence of kriging and maximum entropy estimators. *Mathematical Geology, 29* (1), 131-152.

MATLAB (2006). MATLAB 7.3.0 (R2006b). Natick, MA: The MathWorks.

Orton, T. G. & Lark, R. M. (2007a). Accounting for the uncertainty in the local mean in spatial prediction by Bayesian maximum entropy. *Stochastic Environmental Research Risk Assessment, 21*, 773-784.

Orton, T. G. & Lark, R. M. (2007b). Estimating the local mean for BME by generalized least squares and maximum likelihood, and an application to the spatial analysis of a censored soil variable. *European Journal of Soil Science, 58*, 60-73.

SAS Institute. (2008). *SAS online doc, Version 9.2*. Cary, NC: SAS Institute.

Savelieva, E., Demyanov, V., Kanevski M., Serre, M., & Christakos, G. (2005). BME-based uncertainty assessment of the Chernobyl fallout. *Geoderma, 128* (3-4), 312-324.

Serre, M.L. (1999). Environmental spatiotemporal mapping and ground water flow modelling using the BME and ST methods (Ph.D. Dissertation, University of North Carolina at Chapel Hill, 1999).

Serre, M.L. (2007, July). *Introduction to Bayesian maximum entropy*. Paper presented at the BME workshop sponsored by the Department of Statistics, University of Nebraska-Lincoln.

Serre, M.L. & Christakos, G. (1999). Modern geostatistics: computational BME analysis in the light of uncertain physical knowledge-the Equus Beds study. *Stochastic Environmental Research and Risk Assessment, 13*, 1-26.

Serre, M. L., Kolovos, A., Christakos, G., & Modis, K. (2003). An application of the holistochastic human exposure methodology to naturally occurring arsenic in Bangladesh drinking water. *Risk Analysis, 23* (3), 515-528.

Yu H-L, Kolovos, A., Christakos, G., Chen, J-C, Warmerdam, S. & Dev, B. (2007). Interactive spatiotemporal modelling of health systems: The SEKS-GUI framework. *Stochastic Environmental Research and Risk Assessment-Special Issue on Medical Geography as a Science of Interdisciplinary Knowledge Synthesis under Conditions of Uncertainty*.

Chapter 4 Weighted Kriging vs. Bayesian Maximum Entropy: Triangular

4.1 Introduction

The simulation study in Chapter 3 provided evidence that weighted kriging yields more desirable validation statistics than the BME methodology. However, the soft data came from a Gaussian distribution, and not all soft data will be of this form. This chapter focuses on investigating how the procedures compare when the soft data are of a different form. Proponents of BME claim it shows its strength when the soft data are not symmetric (Kolovos, personal communication, July 8, 2009). Therefore, in this chapter, the soft data were generated using a nonsymmetrical triangular distribution. Another simulation study was used to compare the validation statistics from weighted kriging to those from BME.

4.2 Triangular Soft Data

In the SEKS-GUI, probabilistic soft data can be in the form of a triangular distribution with known mode and upper and lower limits. Thus, each soft data point, say A, consists of its spatial coordinates (xA, yA), lower limit (u1A), mode (mA), and upper limit (u2A) as shown in Figure 4.1 (Kolovos, Yu, & Christakos, 2006).

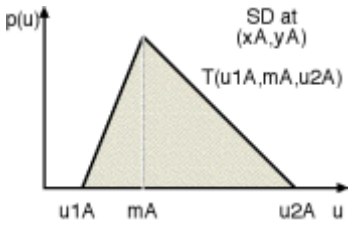

Figure 4.1: Triangular soft data

For this study, the following sequence of events was used to generate soft data by means of the triangular distribution. Recall that a 40 by 40 grid was simulated using a Spherical spatial floor with a specific nugget, range, and sill. Of these 1600 points, 400 were observed values. A certain percentage of these observed values were altered and made into soft data while the remaining observations were unchanged hard data points. To create probabilistic soft data in the form of a triangular distribution, a random number, say x, was sampled from a triangular distribution with a lower limit of 0, upper limit of 5, mode of 1, and mean of 2. These numbers were chosen because it created a positively skewed distribution, and the variance for the stated distribution was 1.167, which is close 1, the sill of the simulated spatial floor. Now suppose one of the observed values which became soft data was denoted by the letter z. The lower limit of the soft data point was defined as u1A = z-(x-0), and the upper limit was u2A = z+(5-x). Thus, the mode, mA, was equal to u1A +1, and the mean was equal to u1A +2. The soft data point was then fully defined by u1A, mA, and u2A.

The soft distribution limits, u1A and u2A, as defined ensured that the observed value z was always within the distribution limits. For example, if x was randomly selected to be the lower limit of 0, then u1A = z-(0-0) = 0 and u2A = z+(5-0) = z+5. In this case, z was contained in the closed interval [z, z+5]. Furthermore, if x was randomly

selected to be the upper limit of 5, then u1A = z-(5-0) = z-5 and u2A = z+(5-5) = z. Again, z was contained in the closed interval [z-5, z]. Likewise, for all other values of x, z was always within the interval limits. Therefore, the proposed transformation identified the value of z, with the randomly selected x, in terms of the relative position of z within the soft distribution limits (Kolovos, personal communication, August 18, 2010).

4.3 Simulation Study

The triangular soft data were generated using SAS® Version 9.2 (SAS Institute, 2008) using the methodology described in Section 4.2. After the hard and soft data files were created, they were used in the weighted kriging procedure in SAS® and then in the SEKS-GUI package in Matlab Version 7.3.0 (2006).

To obtain the BME predictions, the majority of the steps in the SEKS-GUI package were the same as those in Chapter 3. Thus, only those screens which differ are shown in this section. Figures 4.2 and 4.3 are screenshots of the steps involved in selecting the soft data type. Both figures show that the selected type was the triangular distribution.

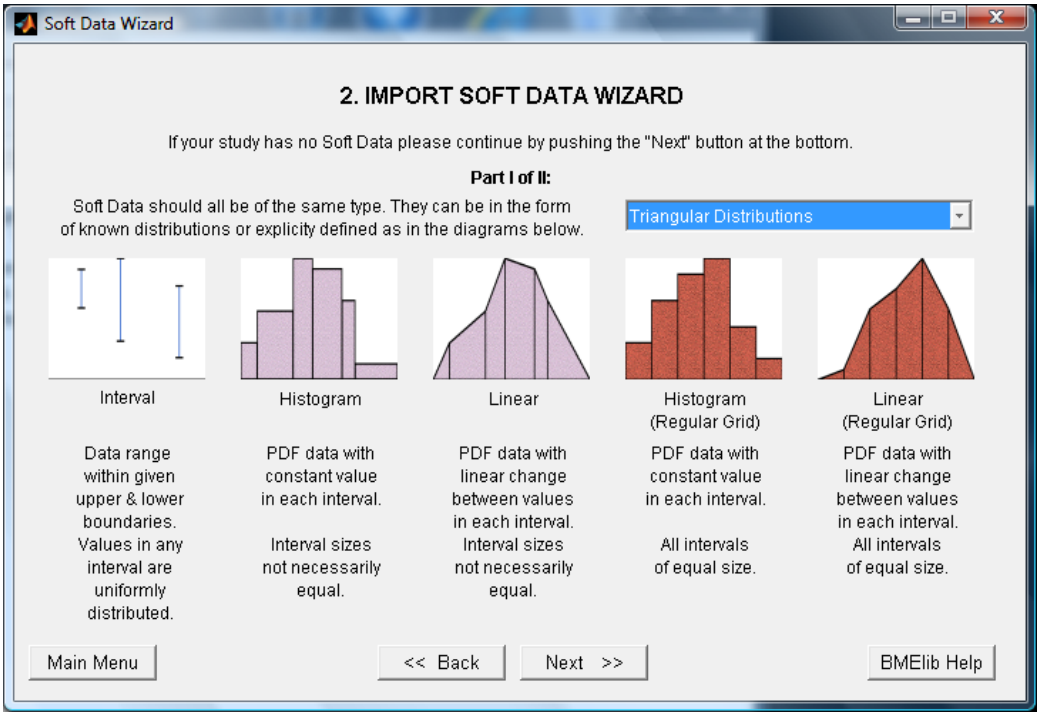

Figure 4.2: A screenshot of the soft data types in SEKS-GUI with Triangular selected

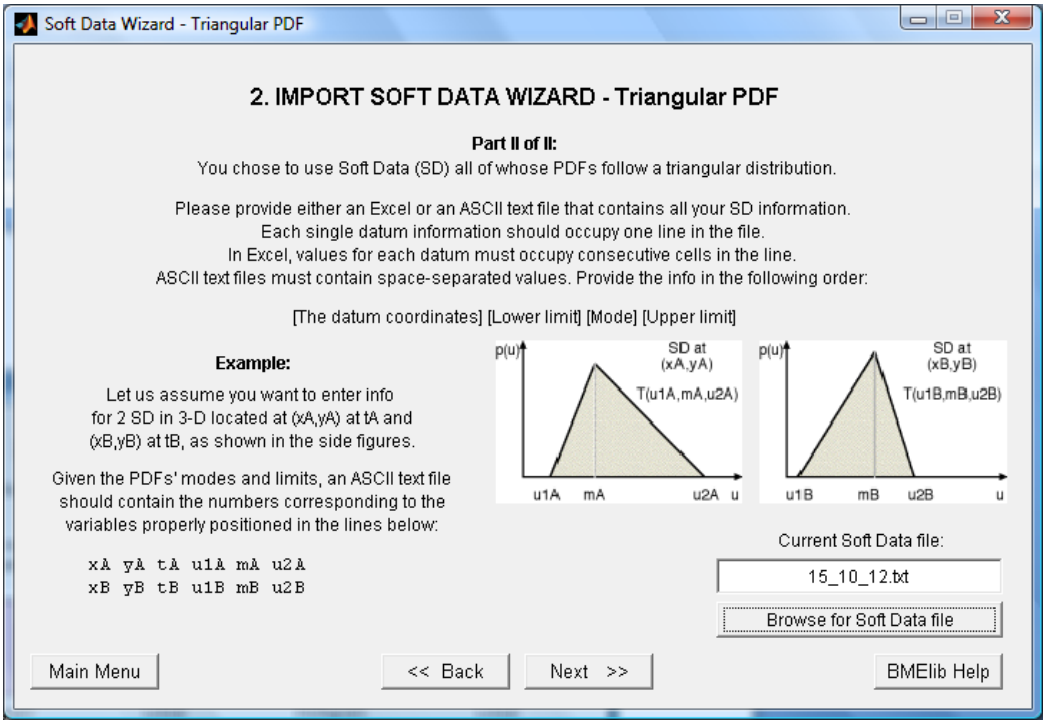

Figure 4.3: A screenshot of importing soft data with Triangular distribution in SEKS-GUI

4.4 Results

The validation statistics defined in Section 2.6 were used to compare the BME approach to the weighted kriging procedure. These two prediction techniques are denoted by **BME** and **Weighted**, respectively, in the tables. In both analyses, the soft data were generated using the triangular distribution described in Section 4.2. Tables 4.1 and 4.2 summarize the means of the validation statistics.

Type	RMSE	AVAR	SME	ABSMPE	RMSSE
BME	0.5919	0.1921	-0.1082	0.4201	1.3917
Weighted	0.3750	0.2563	0.0209	0.2959	0.7626

Table 4.1: Fit statistics obtained from BME and weighted kriging with a range=15 and 10% Triangular soft data

Type	RMSE	AVAR	SME	ABSMPE	RMSSE
BME	0.5492	0.2491	-0.3527	0.4175	1.0325
Weighted	0.3549	0.1429	0.3726	0.2751	0.9852

Table 4.2: Fit statistics obtained from BME and weighted kriging with a range=30 and 50% Triangular soft data

4.5 Conclusions

Friedman's Chi-Square Test was used to determine if there was a significant difference between the prediction procedures. For this nonparametric test, each simulated data set served as a block and the type of prediction was the treatment. For the data sets with a simulation range of 15 and 10% soft data, Friedman's test resulted in

highly significant p-values (<0.0001) for all the statistics listed in the table. For the data sets with a simulation range of 30 and 50% soft data, the RMSSE was the only statistic which did not result in a highly significant p-value. However, for this particular statistic, it is more important to compare it to the desired value of 1 than to compare them to each other. Based on this criterion, the RMSSE from weighted kriging was more desirable. Furthermore, recall that the RMSE should be small for a model which fits the data well, and the SME and ABSMPE should be close to zero. Therefore, with the exception of the SME in Table 4.2, the means from weighted kriging were always more desirable than those from BME. Although the SME in Table 4.2 was an exception, the BME statistic was only 0.0199 closer to 0 than the weighted kriging statistic. Thus, overall, weighted kriging outperformed BME.

4.6 BME Limitations

BME, like ordinary kriging, can be used in estimation and prediction when the local mean is known or assumed to fit some simple model. However, unlike ordinary kriging, in the prior stage of BME, the mean must be calculated from the data alone if the local mean is unknown (Orton & Lark, 2007b). According to Orton and Lark (2007a), when a large number of data points are used to estimate the mean, the uncertainty will be low and the effect on the resulting BME predictions will be minimal. However, hard data are often limited, making it important to incorporate the uncertainty in the prediction.

Furthermore, when the mean is assumed to be given by a constant, linear, or quadratic function, then the parameters are calculated using generalized least squares in

the BMELIB software (Orton & Lark, 2007b). This approach assumes that the mean and variance provide a good representation of the soft data. However, if these parameters do not adequately represent the soft data, this approach can result in errors in the BME predictions. Orton and Lark (2007b) suggest that a maximum likelihood approach produces a better estimate of the local mean if the soft data are of interval form. This approach utilizes the pdf of the soft data and therefore results in more accurate predicted values (Orton & Lark, 2007b).

As mentioned, a generalized least squares approach can lead to inaccurate predictions when the soft data are of the interval form (Orton & Lark, 2007b). According to the SEKS-GUI, this type of soft data is described by an upper and lower boundary, and the values within the interval are uniformly distributed (Kolovos et al., 2006). Since the soft data in the simulation studies in this paper were not of this form, it was appropriate to use the methodology in the BMELIB software.

4.7 References

Kolovos, A., Yu H.-L., & Christakos, G. (2006). *SEKS-GUI v.0.6 user manual*. San Diego: Department of Geography, San Diego State University.

MATLAB (2006). MATLAB 7.3.0 (R2006b). Natick, MA: The MathWorks.

Orton, T. G. & Lark, R. M. (2007a). Accounting for the uncertainty in the local mean in spatial prediction by Bayesian maximum entropy. *Stochastic Environmental Research Risk Assessment, 21*, 773-784.

Orton, T. G. & Lark, R. M. (2007b). Estimating the local mean for BME by generalized least squares and maximum likelihood, and an application to the spatial analysis of a censored soil variable. *European Journal of Soil Science, 58*, 60-73.

SAS Institute. (2008). *SAS online doc, Version 9.2*. Cary, NC: SAS Institute.

Chapter 5 Conclusions

Ordinary kriging is unable to process multiple levels of uncertain information that are often present in environmental studies. This dissertation introduced a spatial prediction technique called weighted kriging to overcome this limitation. The majority of the work was spent on the numerical implementation of this new methodology. Weighted kriging required weight adjustments to estimate the semivariogram parameters and also required adjustments to the semivariogram values used in the kriging matrices. This method was implemented and tested against two alternative kriging procedures. The first alternative used only the hard data in prediction, and the second used both the hard and soft data but treated both as hard. Simulated case studies showed that weighted kriging consistently results in more desirable model fitting statistics.

Prior to this work, Bayesian Maximum Entropy (BME) was the modeling and mapping method often used to incorporate various physical knowledge bases and soft information into spatial analysis. Chapter 3 gave an overview of this approach and the software library used for numerical implementation.

Two simulated case studies were used to compare BME to weighted kriging. The site specific knowledge for the first comparison included hard data and soft data from a Gaussian distribution. It has been shown (Serre, 1999) that when using this type of soft knowledge in combination with hard data, BME yields more desirable results than those from traditional kriging methods. However, the simulated case studies in this work showed that, in terms of validation statistics, weighted kriging outperforms BME.

The site specific knowledge for the second comparison of BME against weighted kriging included hard data and soft data generated from a triangular distribution. Serre (1999) states that "this is a case of considerable interest in spatiotemporal mapping application where uncertain information may be described by means of intervals for the measured attribute." (p. 204). It has been shown (Serre, 1999) that BME provides more accurate predicted values than traditional kriging methods in this situation as well. However, the simulated case studies in this work showed that weighted kriging produces more desirable validation statistics than BME.

An important feature of BME is that when only hard data are used, BME simplifies to ordinary kriging, but when additional sources of knowledge are considered, BME can process this information and produce a more accurate prediction. In other words, "classical geostatistics results are preserved as limited cases of BME analysis" (Serre, 1999, p. 206). Although this is true of BME, it is important to point out that weighted kriging possesses this same quality.

Based on the simulations in this dissertation, the weighted kriging prediction procedure not only possesses considerable flexibility regarding the type of soft data, but it also offers robust prediction. That is, the resulting fit statistics are consistent for different types of soft data and different simulation parameters.

In the future, this research can be extended to not only account for different percentages of uncertain information but also different qualities of uncertain information, represented by differing variances. In the simulation studies in this dissertation, either 10% or 50% of the points were randomly chosen to be soft data. These points became

soft by adding an independent N(0, σ^2) component, where σ^2=0.5. As an extension, multiple values of σ^2 could be investigated. For example, 50% of the soft data could have a 'large' variance with the addition of a N(0, 2.0) component, and the remaining 50% could have a 'small' variance created by the addition of a N(0, 0.5) component.

Furthermore, it may be possible to estimate the softness levels if the data groups are known. If there are two levels of softness, four semivariogram parameters would need to be estimated. Three of these include the nugget, range, and sill of the hard data. The fourth parameter is the additional nugget effect of the soft data. All of these parameters would be estimated by iteratively reweighted least squares, and the semivariogram values would be calculated using the equations in (2.16) and listed again below:

$$\gamma_{HH}(h) = \gamma(h),$$
$$\gamma_{HS}(h) = \gamma(h) + \frac{1}{2}\Delta,$$
$$\gamma_{SS}(h) = \gamma(h) + \Delta.$$

The estimation of the data softness could be also extended from two levels of softness to multiple levels. In terms of the groundwater quality study, consider the nitrate levels from this year, last year, and two years ago. The current nitrate levels could serve as hard data, last year's nitrate levels could serve as the first level of soft data, and the nitrate levels from two years ago could serve as the second level of soft data. Due to the time of collection, last year's data would have less uncertainty than the data collected two years ago, and in order to estimate the softness levels, five semivariogram parameters would need to be estimated. Three of these include the nugget, range, and sill of the hard data. The fourth parameter, Δ_1, is the additional nugget effect corresponding to the first

level of soft data, and the fifth parameter, Δ_2, is the additional nugget effect corresponding to the second level of soft data. All of these parameters would be estimated by iteratively reweighted least squares, and the semivariogram values would be given by:

$$\begin{aligned}
\gamma_{HH}(h) &= \gamma(h), \\
\gamma_{HS_1}(h) &= \gamma(h) + \frac{1}{2}\Delta_1, \\
\gamma_{S_1S_1}(h) &= \gamma(h) + \Delta_1, \\
\gamma_{HS_2}(h) &= \gamma(h) + \frac{1}{2}\Delta_2, \\
\gamma_{S_1S_2}(h) &= \gamma(h) + \frac{1}{2}\Delta_1 + \frac{1}{2}\Delta_2, \\
\gamma_{S_2S_2}(h) &= \gamma(h) + \Delta_2.
\end{aligned}$$

In addition to investigating different qualities of uncertain information, other types of soft data could be generated for use in a simulation study. In this paper, Gaussian and Triangular soft data were used in simulation studies, but the SEKS-GUI package allows for several other types. These include soft data whose probability distributions functions (pdfs) are uniformly distributed, interval soft data where values in any interval are uniformly distributed, histogram soft data, and linear soft data (Kolovos, Yu, & Christakos, 2006). Histogram soft data are data with a constant value in each interval, either on a grid where bins do not necessarily have the same size or on a grid where bins do have the same size (See Figure 5.1). Alternatively, linear soft data are data with a linear change between values in each interval. This type of data can also either be on a grid where bins do not have the same size or on a grid where bins do have the same size (See Figure 5.2).

Figure 5.1: Histogram soft data-interval sizes not equal (left), intervals of equal size (right)

Figure 5.2: Linear soft data-interval sizes not equal (left), intervals of equal size (right)

This research could also be extended to a spatiotemporal analysis rather than a spatial-only investigation. As mentioned in Chapter 3, users of the SEKS-GUI package must specify if the study is purely spatial or if it includes a time variable. If a spatiotemporal analysis is requested, it has the ability to generate predictions at specified points in space and time. Furthermore, future developments could be made by expanding the use of soft data in the area of prediction with covariates, i.e., cokriging. Simulation studies could be used to investigate different percentages and different qualities of uncertain information in the primary variable in combination with different percentages and qualities of the covariate.

Bibliography

Choi, K. M., Serre, M. L., & Christakos, G. (2003). Efficient mapping of California mortality fields at different spatial scales. *Journal of Exposure Analysis and Environmental Epidemiology, 13* (2), 120-133.

Christakos, G. (1990). A Bayesian/maximum-entropy view to the spatial estimation problem. *Mathematical Geology, 22* (7), 763-777.

Christakos, G. (2009). *Epistematics: An evolutionary framework of real world problem-solving*. New York: Springer.

Christakos, G. & Li, X. (1998). Bayesian maximum entropy analysis and mapping: A farewell to kriging estimators? *Mathematical Geology, 30* (4), 435-462.

Clark, I. & Harper, W. V. (2000). *Practical geostatistics*. Columbus, Ohio: Ecosse North America.

Cressie, N. (1991). *Statistics for spatial data*. New York: John Wiley & Sons.

ESRI. (2001). *Using ArcGIS geostatistical analyst*. Redlands, CA: ESRI.

ESRI. (2006). *ArcGIS desktop help 9.2*. Redlands, CA: ESRI.

Isaaks, E. H. & Srivastava, R. M. (1989). *Applied geostatistics.* New York: Oxford University Press.

Journel, A. G. (1986). Constrained interpolation and qualitative information-The soft kriging approach. *Mathematical Geology, 18* (3), 269-286.

Journel, A. G. & Huijbregts, Ch. J. (1978). *Mining geostatistics*. New York: Academic Press.

Kolovos, A. (2001). Computational investigations of the BME mapping approach and incorporation of physical knowledge bases (Ph.D. Dissertation. University of North Carolina at Chapel Hill, 2001).

Kolovos, A., Yu H.-L., & Christakos, G. (2006). *SEKS-GUI v.0.6 user manual.* San Diego: Department of Geography, San Diego State University.

Law, D. C. G., Bernstein, K. T., Serre, M. L., Schumacher, C. M., Leone P. A., Zenilman, J. M., et al. (2006). Modeling a syphilis outbreak through space and time using the Bayesian maximum entropy approach. *Annals of Epidemiology, 16* (11), 797-804.

Lee, Y.-M., & Ellis, J. H. (1997). On the equivalence of kriging and maximum entropy estimators. *Mathematical Geology, 29* (1), 131-152.

Liedtke, M., Marx D., & Kachman S. (2009, January). Incorporating soft data in the kriging equations. Paper presented at the 8th Annual Hawaii International Conference on Statistics, Mathematics and Related Mathematics, Honolulu, Hawaii.

Matheron, G. (1963). Principles of geostatistics. *Economic Geology and the Bulletin of the Society of Economic Geologists*, 58, 1246-1266.

MATLAB (2006). MATLAB 7.3.0 (R2006b). Natick, MA: The MathWorks.

Olea, R. A. (2006). A six-step practical approach to semivariogram modeling. *Stochastic Environmental Research Risk Assessment, 20*, 307-318.

Orton, T. G. & Lark, R. M. (2007a). Accounting for the uncertainty in the local mean in spatial prediction by Bayesian maximum entropy. *Stochastic Environmental Research Risk Assessment, 21*, 773-784.

Orton, T. G. & Lark, R. M. (2007b). Estimating the local mean for BME by generalized least squares and maximum likelihood, and an application to the spatial analysis of a censored soil variable. *European Journal of Soil Science, 58*, 60-73.

SAS Institute. (2008). *SAS online doc, Version 9.2*. Cary, NC: SAS Institute.

Savelieva, E., Demyanov, V., Kanevski M., Serre, M., & Christakos, G. (2005). BME-based uncertainty assessment of the Chernobyl fallout. *Geoderma, 128* (3-4), 312-324.

Serre, M.L. (1999). Environmental spatiotemporal mapping and ground water flow modelling using the BME and ST methods (Ph.D. Dissertation, University of North Carolina at Chapel Hill, 1999).

Serre, M.L. (2007, July). *Introduction to Bayesian maximum entropy*. Paper presented at the BME workshop sponsored by the Department of Statistics, University of Nebraska-Lincoln.

Serre, M.L. & Christakos, G. (1999). Modern geostatistics: computational BME analysis in the light of uncertain physical knowledge-the Equus Beds study. *Stochastic Environmental Research and Risk Assessment, 13*, 1-26.

Serre, M. L., Kolovos, A., Christakos, G., & Modis, K. (2003). An application of the holistochastic human exposure methodology to naturally occurring arsenic in Bangladesh drinking water. *Risk Analysis, 23* (3), 515-528.

Schabenberger, O. & Gotway, C. A. (2005) *Statistical methods for spatial data analysis*. Boca Raton, Florida: Chapman & Hall/CRC.

Yu H-L, Kolovos, A., Christakos, G., Chen, J-C, Warmerdam, S. & Dev, B. (2007). Interactive spatiotemporal modelling of health systems: The SEKS-GUI framework. *Stochastic Environmental Research and Risk Assessment-Special Issue on Medical Geography as a Science of Interdisciplinary Knowledge Synthesis under Conditions of Uncertainty*.

APPENDIX A

SAS® code for weighted kriging with a range=15 and 10% soft data from Triangular distribution

```
libname skew "C:\skewed";

ODS RESULTS OFF;
ods listing close;

*** START A MACRO TO ITERATE PROCESS **;
%LET CASE = RW15_10_0;
%LET RANGE = 15;
%LET ITERATIONS=3;
%LET Q=1;

TITLE " &CASE, RANGE = &RANGE, &ITERATIONS ITERATIONS";

Data PARMSCOVSRW&Q;  Set _NULL_;
Data STATRW&Q;  Set _NULL_;
Data savedata;  Set _NULL_;
Data predval;  Set _NULL_;

Data PARMSCOVSHS&Q;  SET _NULL_;
Data STATHS&Q;  SET _NULL_;
Data PARMSCOVSH&Q; SET _NULL_;
Data STATH&Q; SET _NULL_;
Data SKEWSOFT&Q; SET _NULL_;

%MACRO sp4040 ( n4040 );    * - start macro P -;

%DO I=1 %TO &n4040;  * - set # of iterations -;

%Let seed1=3043248+&I&Q;
%Let seed2=3089723+&I&Q;
%Let seed3=3061258+&I&Q;

*** GENERATE SPHERICAL SPATIAL FLOOR **;
DATA A;
DO  LAT = 1 TO 40;
   DO  LNG = 1 TO 40;
     OUTPUT;
   END;
 END;
RUN;
PROC IML WORKSIZE=320;
USE A;
READ ALL;


NUGGET=00;
RANGE=15; *&RANGE;  * CHANGE RANGE HERE  ;
SILL=1;
```

```
NOBS= NROW(LAT);
H=J(NOBS,NOBS,.);

START;

DO I=1 TO NOBS;
 DO J=I TO NOBS;
  H(|I,J|)=
    SQRT(((LAT(|I,1|)-LAT(|J,1|))##2)+((LNG(|I,1|)-LNG(|J,1|))##2));
  H(|J,I|)=H(|I,J|);
 END;
END;

H=H><RANGE;

A1=H#(-1.5)#(SILL/RANGE);
A2=(H##3)#0.5#(SILL/(RANGE##3));
FREE H;
A3=J(NOBS,NOBS,SILL);
A0=I(NOBS)#NUGGET;
A4=A1+A2+A3+A0;
A4 = ROOT(A4);
FREE A0 A1 A2 A3;

E=J(NOBS,1,.);
DO I=1 TO NOBS;
 E(|I,1|)=RANNOR(&seed1);
END;

E=1#E;
Y=A4`*E+10;      * Y is the spatial floor*;

SPH2 = LAT||LNG||Y;
*PRINT SPH2;
COLS='LAT'||'LNG'||'Y';
CREATE DATOUT FROM SPH2(|COLNAME=COLS|);
APPEND FROM SPH2;
FINISH;
RUN;

DATA OUT;
  SET DATOUT;
  FILE 'OUT400.DAT ';
  PUT LAT LNG Y;
  Run;

data obs;
      do i=1 to 1600;
      x=ranuni(&seed2);
      output;
   end;
run;
```

```
data combine;
  merge obs out;
run;

proc sort data=combine;
  by x;
run;

data id;
   do id=1 to 1600;
   output;
  end;
run;

data combine2;
  merge id combine;
run;

*** 10% OR 40 RANDOM OBSERVATIONS ARE MADE SOFT **;
data soft;
  set combine2;
  if id<41 then
    do;
         triang=rantri(0,.2)*5;  *** SAMPLE FROM TRIANGULAR
            DISTRIBUTION WITH LOWER BOUND=0, UPPER BOUND=5, MODE=1 **;
         Ylow=Y-triang;
         Ym=1+Ylow;
         Yup=Y+(5-triang);
      end;
      output;
run;

data onlysoft;
  set soft;
  keep lat lng Ylow Ym Yup;
  if id<41;
run;

data observed;
  set soft;
      if id<401;
      if id<41 then type=1;  *Type=1=soft data  Type=0=hard data;
       else type=0;
run;

data observed;
  set observed;
  if type=1 then Y=Ym;
run;

data validation;
  set soft;
    if id>400;
run;
```

```
data hard;
  set soft;
  if id>40 and id<401;;
run;

proc sort data=observed;
  by lat lng;
run;

proc variogram data=observed outpair=z;
  var Y;
  coordinates xc=lat yc=lng;
  compute novariogram;
run;

proc variogram data=observed outpair=h;
  var type;
  coordinates xc=lat yc=lng;
  compute novariogram;
run;

data h;
  set h;
  h1=v1; h2=v2;
  h3=h1+h2;
  drop v1 v2 distance cos varname;
run;

proc sort data=h;
  by x1 y1 x2 y2;
run;
proc sort data=z;
  by x1 y1 x2 y2;
run;

data pair; merge z h;
  by x1 y1 x2 y2;
  variog=(v1-v2)**2;
run;

*** OBTAIN SEMIVARIOGRAM ESTIMATES **;
PROC NLIN DATA=pair METHOD=NEWTON NOHALVE;
  TITLE 'SPHERICAL MODEL';
  PARMS N=0,.25,  S=.25,.5,1,1.5 R=14,15,15.5,16;
 Q1 = 1.5*Distance/R;
 Q2 = .5*Distance**3/R**3;
  IF Distance < R and h3=0 THEN DO;
    _WEIGHT_=1/sqrt(S);
    MODEL variog = S*(Q1-Q2);
  END;
  ELSE IF DISTANCE>=R AND H3=0 THEN DO;
     _WEIGHT_=1/sqrt(S);
    MODEL variog = S;
  END;
```

```
  ELSE IF Distance < R AND H3=1 THEN DO;
    _WEIGHT_=1/(sqrt(S + .5*N));
    MODEL variog = .5*N + S*(Q1-Q2);
  END;
  ELSE IF DISTANCE>=R AND H3=1 THEN DO;
    _WEIGHT_=1/(sqrt(S + .5*N));
    MODEL variog = .5*N+ S;
  END;
  ELSE IF Distance < R AND H3=2 THEN DO;
    _WEIGHT_=1/(sqrt(S + N));
    MODEL variog = N + S*(Q1-Q2);
  END;
  ELSE IF DISTANCE>=R AND H3=2 THEN DO;
    _WEIGHT_=1/(sqrt(S + N));
    MODEL variog = N + S;
  END;
  ods output 'Parameter Summary'=parm;
  ODS OUTPUT 'Convergence Status'=CVSTAT;
  ods output 'Summary Statistics : Dependent Variable VARIOG'=anova;
RUN;

data anova1;
  set anova;
  keep ms;
run;

proc transpose data=anova1 out=anova2;
run;

data anova2;
  set anova2;
  nlinmse=COL2;
  keep nlinmse;
run;

data parm1;
  set parm;
  keep estimate;
run;

proc transpose data=parm1 out=parm2;
  run;

data parms;
  set parm2;
  varname="Y";
  nugget=COL1;
  scale=COL2;
  range=COL3;
  form="SPH";
  keep nugget range scale form varname;
run;
```

```
*********************************************;
*       CHECKING QUALITY OF PARAMETERS       ;
*********************************************;
DATA CHECK; * 1 = GOOD , 2=BAD;
  MERGE PARMS CVSTAT;
  IF RANGE<1   | RANGE>50 THEN PSTATUS=2; ELSE
  IF SCALE<0.1 | SCALE >20 THEN PSTATUS=2; ELSE PSTATUS=1;

  IF Reason = 'NOTE: Convergence criterion met.' & PSTATUS=1 THEN
CHECKVAL=1; ELSE
  IF Reason = 'NOTE: Convergence criterion met.' & PSTATUS=2 THEN
CHECKVAL=2; ELSE CHECKVAL=3;
  CALL SYMPUT ('CHECKVAL', CHECKVAL);
  KEEP CHECKVAL;
RUN;

%PUT CHECKVAL=&CHECKVAL;
RUN;

*Change parameters if needed*;
data converge;
  merge check parms;
  if checkval=2 or checkval=3 then
  do;
   nugget=.5;
   scale=1.25;
   range=15;
  end;
run;

data parms;
  set converge;
  drop checkval;
run;

DATA COVPARMS; SET PARMS;
  ITER=&i;
  CHECK=&CHECKVAL;
RUN;

%let npoints=20;

proc krige2d data=observed oute=est outn=near;
  coord xc=lat yc=lng;
  *grid x=1 to 40 y=1 to 40;
  predict npoints=&npoints var=Y;
  grid gdata=soft xcoord=lat ycoord=lng;
  model mdata=parms;
run;

data semi;
  merge near parms;
  by varname;
run;
```

```
data semi; set semi;
  x1=gxc; y1=gyc; x2=xc; y2=yc;
  drop gxc gyc xc yc form label;
run;

data identify; set semi;
  obsnum=_N_;
  sample=int((obsnum-1)/20)+1;
run;

%LET LOOP=1600;

%MACRO pred20 ( n );     * - start macro P -;

%DO L=1 %TO &n;          * - set # of iterations -;

****START LOOP TO PICK 1 POINT AT A TIME TO BE PREDICTED****;
data predict;
  set identify;
  if sample=&L;
run;

proc sort data=predict;
  by x2 y2;
run;

proc variogram data=predict outpair=pred;
  var value;
  coordinates xc=x2 yc=y2;
  compute novariogram;
run;

data pred; set pred;
  N=1;
run;
proc sort data=pred;
  by x1 y1 x2 y2;
run;
data variog;
  merge pair pred;
  by x1 y1 x2 y2;
  if N=1;
  keep x1 y1 x2 y2 h1 h2 h3 distance;
run;

data variog;
  set variog;
  varname="Y";
run;
```

```
*** CALCULATE ORDINARY KRIGING MATRIX C VALUES **;
data variogram;
  merge parms variog;
  by varname;
  dist=distance;
  if dist=0 then varioc=0;
  else if dist < range & h3=0 then varioc= scale*(1.5*dist/range-
      .5*dist**3/range**3);
  else if dist < range & h3=1 then varioc=scale*(1.5*dist/range-
      .5*dist**3/range**3)+ .5*nugget;
  else if dist < range & h3=2 then varioc=scale*(1.5*dist/range-
      .5*dist**3/range**3)+ nugget;
  else if dist > range & h3=0 then varioc=scale;
  else if dist > range & h3=1 then varioc=scale+.5*nugget;
  else if dist > range & h3=2 then varioc=scale + nugget;
  else varioc=.;
run;


data predict1;
  set predict;
  distance=sqrt((x1-x2)**2+(y1-y2)**2);
  lat=x2; lng=y2;
  drop x2 y2;
run;

data d;
  merge predict1 observed;
  by lat lng;
  if value=. then delete;
run;

*** CALCULATE ORDINARY KRIGING MATRIX D **;
data vard;
  set d;
  dist=distance;
  if dist=0 then variod=0;
  else if dist < range & type=0 then variod= scale*(1.5*dist/range-
      .5*dist**3/range**3);
  else if dist < range & type=1 then variod= scale*(1.5*dist/range-
      .5*dist**3/range**3)+ .5*nugget;
  else if dist > range & type=0 then variod=scale;
  else if dist > range & type=1 then variod=scale+.5*nugget;
  else variod=.;
run;

data observed2;
  set observed;
  x1=lat; y1=lng;
  drop lat lng;
run;
```

```
data d2;
  merge predict1 observed2;
  by x1 y1;
  if value=. then delete;
run;

*** CONSTRUCT MATRIX C **;

%let npoints=20;
proc iml;
  use variogram;
  read all;

  c1 = j(&npoints,&npoints,0);
  k=0;
  do i = 1 to &npoints-1;
    do j = i+1 to &npoints;
        k = k+1;
        c1[i,j] = varioc[k];
        c1[j,i] = c1[i,j];
      end;
end;

print c1;

  use vard;
  read all;

  diag=j(&npoints,&npoints,0);
  k=0;
  do i=1 to &npoints;
      k=k+1;
  diag[i,i]=0;
end;

  x1=c1+diag;

  jend=j(&npoints,1,1);
  jrow=j(1,&npoints,1);
  jdot=j(1,1,0);
  x2=x1||jend;
  jr1=jrow||jdot;
  c=x2//jr1;
print c1 diag x1 c x2;

*** CONSTRUCT MATRIX D **;
  xd = j(&npoints,1,0);
  val = j(&npoints,1,0);
  k=0;
  do i = 1 to &npoints;
      k = k+1;
        xd[i] = variod[k];
        val[i]= value[k];
  end;
```

```
  jdot=j(1,1,1);
  d=xd//jdot;

print xd d val;

** CALCULATE WEIGHTS**;
w=inv(C)*D;
w1=w[1:&npoints];
** CALCULATE PREDICTED VALUE**;
predict=w1`*val;
check=w1`*j(&npoints,1,1);
scale=scale[1];
nugget=nugget[1];
var1=scale+.5*nugget;

use d2;
read all;

x=x1[1];  y=y1[1];

*** CALCULATE PREDICTION VARIANCE**;
 if type[1]=1 then var=var1-w`*D; else var=w`*D;
 var=var[1];

print w w1 predict check var x y;

pred2 = predict||var||x||y;

COLS='pred'||'var'||'x'||'y';
CREATE DATOUT2 FROM pred2(|COLNAME=COLS|);
APPEND FROM pred2;
RUN;

DM log 'clear';
DM output 'clear';

data predval; set predval datout2;
  if x;
run;

%END;       * MAIN MACRO ENDS ITERATIONS;

%MEND pred20;

*--------------------;
%pred20 (&LOOP);

**************************;

data predval; set predval;
  lat=x; lng=y;
  drop x y;
run;
```

```
proc sort data=predval;
  by lat lng;
run;
proc sort data=soft;
  by lat lng;
run;

data comparison;
  merge predval soft;
  by lat lng;
  stderr=sqrt(var);
run;

proc sort data=comparison;
  by var;
run;

data valid;
  set comparison;
  if id>400;
run;

***************************;
** COMPUTE FIT STATISTICS **;
data compare;
  set valid;
  resid=(Y-pred);
  residsq=(Y-pred)**2;
  absresid=abs(Y-pred);
  msenum=(Y-pred)/stderr;
  msenumsq=((Y-pred)/stderr)**2;
run;

proc univariate data=compare;
  output out=summary sum=mpe rmse ase abmpe avar mse msesq;
  var resid residsq stderr absresid var msenum msenumsq;
run;

data statistics;
 set summary;
 mpe=mpe/1200;
 absmpe=abmpe/1200;
 rmse=sqrt(rmse/1200);
 ase=ase/1200;
 avar=avar/1200;
 mse=mse/1200;
 rmsse=sqrt(msesq/1200);
 run;

data allstats;
  merge statistics anova2 covparms;
run;
```

```
DATA statrw&Q; SET allstats statrw&Q;
RUN;

DATA PARMSCOVSRW&Q; SET  PARMSCOVSRW&Q COVPARMS;
Run;

DATA SKEWSOFT&Q;  SET SKEWSOFT&Q onlysoft;
RUN;

**********************************
*Hard and soft data treated as hard*;

proc variogram data=observed outvar=v;
  var Y;
  coordinates xc=lat yc=lng;
  compute lagd=2 nd=1 maxlag=50;
run;

PROC NLIN DATA=v METHOD=NEWTON; *converge=.01;
  TITLE 'SPHERICAL MODEL';
  _WEIGHT_ = COUNT;
 PARMS N=0,.25,  S=.25,.5,1,1.5 R=14,15,15.5,16;
 Q1 = 1.5*Distance/R;
 Q2 = .5*Distance**3/R**3;

  IF Distance < R THEN DO;
    MODEL variog =  S*(Q1-Q2);
  END;
  ELSE DO;
    MODEL variog =  S;
  END;
  ods output 'Parameter Summary'=parm;
  ODS OUTPUT 'Convergence Status'=CVSTAT;
  ods output 'Summary Statistics : Dependent Variable VARIOG'=anova;
RUN;

data anova1;
  set anova;
  keep ms;
run;

proc transpose data=anova1 out=anova2;
run;

data anova2;
  set anova2;
  nlinmse=COL2;
  keep nlinmse;
run;

data parm1;
  set parm;
  keep estimate;
run;
```

```
proc transpose data=parm1 out=parm2;
  run;

data parms;
  set parm2;
  nugget=COL1;
  scale=COL2;
  range=COL3;
  form="SPH";
  keep nugget range scale form;
run;

*********************************************;
*       CHECKING QUALITY OF PARAMETERS      ;
*********************************************;

DATA CHECK;        * 1 = GOOD , 2=BAD;
  MERGE PARMS CVSTAT;
  IF RANGE<1   | RANGE>50 THEN PSTATUS=2; ELSE
  IF SCALE<0.1 | SCALE >20 THEN PSTATUS=2; ELSE PSTATUS=1;
  IF Reason = 'NOTE: Convergence criterion met.' & PSTATUS=1 THEN
      CHECKVAL=1; ELSE
  IF Reason = 'NOTE: Convergence criterion met.' & PSTATUS=2 THEN
      CHECKVAL=2; ELSE CHECKVAL=3;
  CALL SYMPUT ('CHECKVAL', CHECKVAL);
  KEEP CHECKVAL;
RUN;

%PUT CHECKVAL=&CHECKVAL;
RUN;

data converge;
  merge check parms;
  if checkval=2 or checkval=3 then
  do;
   nugget=0;
   scale=1;
   range=15;
  end;
run;
data parms;
  set converge;
  drop checkval;
run;
************;
DATA COVPARMS; SET PARMS;
  ITER=&i;
  CHECK=&CHECKVAL;
RUN;

proc krige2d data=observed oute=est;
  coord xc=lat yc=lng;
  predict npoints=20 var=Y;
```

```
  grid gdata=validation xcoord=lat ycoord=lng;
  model mdata=parms;
run;

data est2;
  set est;
  lat=gxc;
  lng=gyc;
  keep lat lng estimate stderr;
run;

proc sort data=validation;
  by lat lng;
run;

proc sort data=est2;
  by lat lng;
run;

data compare;
  merge est2 validation;
  by lat lng;
run;

data compare;
  set compare;
  resid=(Y-estimate);
  residsq=(Y-estimate)**2;
  absresid=abs(Y-estimate);
  var=stderr**2;
  msenum=(Y-estimate)/stderr;
  msenumsq=((Y-estimate)/stderr)**2;
run;

proc univariate data=compare;
  output out=summary sum=mpe rmse ase abmpe avar mse msesq;
  var resid residsq stderr absresid var msenum msenumsq;
run;

data statistics;
 set summary;
 mpe=mpe/1200;
 absmpe=abmpe/1200;
 rmse=sqrt(rmse/1200);
 ase=ase/1200;
 avar=avar/1200;
 mse=mse/1200;
 rmsse=sqrt(msesq/1200);
 run;

data allstats;
  merge statistics anova2 covparms;
run;
```

```
DATA STATHS&Q; SET allstats STATHS&Q;
RUN;

DATA PARMSCOVSHS&Q; SET  PARMSCOVSHS&Q COVPARMS;
Run;


*End of hard and soft treated as hard;
************************************************;
*Start of using only hard data;

proc variogram data=hard outvar=v;
  var Y;
  coordinates xc=lat yc=lng;
  compute lagd=2 nd=1 maxlag=50;
run;


PROC NLIN DATA=v METHOD=NEWTON; *converge=.01;
  TITLE 'SPHERICAL MODEL';
  _WEIGHT_ = COUNT;
 PARMS N=0,.25,  S=.25,.5,1,1.5 R=14,15,15.5,16;
 Q1 = 1.5*Distance/R;
 Q2 = .5*Distance**3/R**3;

  IF Distance < R THEN DO;
    MODEL variog =  S*(Q1-Q2);
  END;
  ELSE DO;
    MODEL variog =  S;
  END;
  ods output 'Parameter Summary'=parm;
  ODS OUTPUT 'Convergence Status'=CVSTAT;
  ods output 'Summary Statistics : Dependent Variable VARIOG'=anova;
RUN;

data anova1;
  set anova;
  keep ms;
run;

proc transpose data=anova1 out=anova2;
run;

data anova2;
  set anova2;
  nlinmse=COL2;
  keep nlinmse;
run;

data parm1;
  set parm;
  keep estimate;
run;
```

```
proc transpose data=parm1 out=parm2;
  run;

data parms;
  set parm2;
  nugget=COL1;
  scale=COL2;
  range=COL3;
  form="SPH";
  keep nugget range scale form;
run;

*********************************************;
*       CHECKING QUALITY OF PARAMETERS      ;
*********************************************;

DATA CHECK;        * 1 = GOOD , 2=BAD;
  MERGE PARMS CVSTAT;
  IF RANGE<1   | RANGE>50 THEN PSTATUS=2; ELSE
  IF SCALE<0.1 | SCALE >20 THEN PSTATUS=2; ELSE PSTATUS=1;
  IF Reason = 'NOTE: Convergence criterion met.' & PSTATUS=1 THEN
CHECKVAL=1; ELSE
  IF Reason = 'NOTE: Convergence criterion met.' & PSTATUS=2 THEN
CHECKVAL=2; ELSE CHECKVAL=3;
  CALL SYMPUT ('CHECKVAL', CHECKVAL);
  KEEP CHECKVAL;
RUN;

%PUT CHECKVAL=&CHECKVAL;
RUN;

data converge;
  merge check parms;
  if checkval=2 or checkval=3 then
  do;
   nugget=0;
   scale=1;
   range=15;
  end;
run;
data parms;
  set converge;
  drop checkval;
run;

DATA COVPARMS; SET PARMS;
  ITER=&i;
  CHECK=&CHECKVAL;
RUN;

proc krige2d data=observed oute=est;
  coord xc=lat yc=lng;
  *grid x=1 to 40 y=1 to 40;
```

```
  predict npoints=20 var=Y;
  grid gdata=validation xcoord=lat ycoord=lng;
  model mdata=parms;
run;

data est2;
  set est;
  lat=gxc;
  lng=gyc;
  keep lat lng estimate stderr;
run;

proc sort data=validation;
  by lat lng;
run;

proc sort data=est2;
  by lat lng;
run;

data compare;
  merge est2 validation;
  by lat lng;
run;

data compare;
  set compare;
  resid=(Y-estimate);
  residsq=(Y-estimate)**2;
  absresid=abs(Y-estimate);
  var=stderr**2;
  msenum=(Y-estimate)/stderr;
  msenumsq=((Y-estimate)/stderr)**2;
run;

proc univariate data=compare;
  output out=summary sum=mpe rmse ase abmpe avar mse msesq;
  var resid residsq stderr absresid var msenum msenumsq;
run;

data statistics;
 set summary;
 mpe=mpe/1200;
 absmpe=abmpe/1200;
 rmse=sqrt(rmse/1200);
 ase=ase/1200;
 avar=avar/1200;
 mse=mse/1200;
 rmsse=sqrt(msesq/1200);
 run;

data allstats;
  merge statistics anova2 covparms;
run;
```

```
DATA STATH&Q; SET allstats STATH&Q;
RUN;

DATA PARMSCOVSH&Q; SET  PARMSCOVSH&Q COVPARMS;
Run;
*End of hard data only;

*********************************************************;
*  CLEAR LOG AND OUTPUT WINDOW AFTER EACH ITERATION      ;
*********************************************************;

DM log 'clear';
DM output 'clear';

%END;                        * MAIN MACRO ENDS ITERATIONS;

%MEND sp4040;

*--------------------;
%sp4040 (&ITERATIONS);
*--------------------;
*********************************************************;
* WRITE STATISTICS, PARAMETERS, AND DATA TO EXTERNAL FILES;
*********************************************************;

DATA SKEW.&CASE.rw&Q; SET statrw&Q;
 RUN;
DATA SKEW.&CASE.SOFT&Q;  SET SKEWSOFT&Q;
 RUN;
```

APPENDIX B

The following changes need to made to the SAS® code in Appendix A for the weighted kriging case with a range=15 and 10% Gaussian soft data

```
data soft;
  set combine2;
  if id<41 then
    do;
         Y=Y+rannor(&seed3)*sqrt(.5);
      end;
      output;
run;
```

APPENDIX C

The following changes need to made to the SAS® code in Appendix A for the weighted kriging case with a range=30 and 50% Triangular soft data

```
%LET RANGE = 30;

RANGE = 30;

data soft;
  set combine2;
  if id<201 then
    do;
         triang=rantri(0,.2)*5;
         Ylow=Y-triang;
         Ym=1+Ylow;
         Yup=Y+(5-triang);
      end;
      output;
run;

data onlysoft;
  set soft;
  keep lat lng Ylow Ym Yup;
  if id<201;
run;

data observed;
  set soft;
      if id<401;
      if id<201 then type=1;  *Type=1=soft data  Type=0=hard data;
       else type=0;
run;
```

```
data observed;
  set observed;
  if type=1 then Y=Ym;
run;

data validation;
  set soft;
    if id>400;
run;

data hard;
  set soft;
  if id>200 and id<401;;
run;
```

APPENDIX D

The following changes need to made to the SAS® code in Appendix C for the weighted kriging case with a range of 30 and 50% Gausian soft data

```
data soft;
  set combine2;
  if id<201 then
    do;
         Y=Y+rannor(&seed3)*sqrt(.5);
      end;
      output;
run;
```